DOCUMENT RESUME

ED 480 705 CE 085 471

TITLE	Evaluating the Net Impact of School-to-Work: Proceedings of a Roundtable.
INSTITUTION	National School-to-Work Opportunities Office, Washington, DC.; Department of Education, Washington, DC.; Department of Labor, Washington, DC.
PUB DATE	1997-00-00
NOTE	289p.
AVAILABLE FROM	For full text: http://wdr.doleta.gov/opr/fulltext/ 97-net_impact.pdf.
PUB TYPE	Reports - Evaluative (142)
EDRS PRICE	EDRS Price MF01/PC12 Plus Postage.
DESCRIPTORS	*Education Work Relationship; Educational Research; *Evaluation Problems; *Evaluation Research; *Outcomes of Education; *Policy Analysis; Research Design; Research Methodology; Research Problems; Secondary Education; Strategic Planning; Transitional Programs; Vocational Education
IDENTIFIERS	*Impact Evaluation; *School to Work Opportunities Act 1994

ABSTRACT

This volume brings together the reports and discussion connected with a roundtable convened by the Departments of Education and Labor and the National School-to-Work Office to discuss issues surrounding the conduct of a net impact evaluation of school-to-work. After an introduction that summarizes the approach and intentions of the School-to-Work Opportunities Act and describes evaluation activities that have been commissioned separately from a net impact effort, six commissioned papers are presented. The papers are as follows: "Evaluating Early Program Experiences in the School-to-Work Opportunities Act: Policy and Design Issues" (Gary Burtless); "Net Impact of School-to-Work: Exploring Alternatives" (Charles Dayton); "Net Impact Evaluation of School-to-Work: Desirable but Feasible?" (Robert Glover, Christopher T. King); "Evaluation of School-to-Work Transitional Programs" (James J. Heckman); "Evaluating the School-to-Work Opportunities Act of 1994" (Robert A. Moffitt); and "Net Impact Evaluation of School-to-Work: Contending Expectations" (Hillard Pouncy, Robinson Hollister). Following the papers is a summary of the roundtable discussion and an epilogue that synthesizes both the papers and the discussion. Contains a list of participants, 190 references, and some papers includes tables/figures. (MO)

Reproductions supplied by EDRS are the best that can be made
from the original document.

Evaluating the Net Impact of School-to-Work: Proceedings of a Roundtable

U.S. Department of Labor
Robert B. Reich, Secretary

Employment and Training Administration
Timothy Barnicle, Assistant Secretary

Office of Policy and Research
Gerard F. Fiala, Administrator

1997

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ED 480 705

CE085471

BEST COPY AVAILABLE

Table of Contents

Page Number

Preface 3

List of Participants 5

Introduction: School-to-Work And School-to-Work Evaluation and Research
I. Introduction 7
II. The School-to-Work Opportunities Act of 1994 7
III. School-to-Work Evaluation 12
IV. Conclusion.. 14

Evaluating Early Program Experiences in the School-to-Work Opportunities Act:
Policy and Design Issues
by Gary Burtless
I. Introduction and Summary 19
II. Conceptual and Policy Issues 21
III. Design Issues 30
IV. An Overall Design for the National Evaluation . . . 38
V. Summary 47

Net Impact of School-to-Work: Exploring Alternatives
by Charles Dayton
I. Introduction and Background 49
II. Conceptual and Policy Issues 50
III. Design Issues 62
IV. Implementation Issues 75
V. Conclusions 79

Net Impact Evaluation of School-to-Work: Desirable but Feasible?
by Robert Glover and Christopher T. King
I. Introduction and Background 81
II. Conceptual and Policy Issues 87
III. Design Issues 100
IV. Implementation Issues 115
V. Conclusions 121

Evaluation of School-to-Work Transition Programs
by James J. Heckman

I. Introduction and Summary 125
II. The Method of Instrumental Variables 127
III. Randomized Trials 145
IV. Matching as an Evaluation Estimator 156
Appendix A: A More General Nonparametric Model for Conditional Means 190
Appendix B: Comparison With Conventional Random Coefficient Models 192

Evaluating the School-to-Work Opportunities Act of 1994
by Robert A. Moffitt

I. Introduction and Background 195
II. Conceptual and Policy Issues 198
III. Design Issues 199
IV. Implementation 212
V. Conclusions 212

Net Impact Evaluation of School-to-Work: Contending Expectations
by Hillard Pouncy and Robinson Hollister

I. Introduction and Background 215
II. Conceptual and Policy Issues 216
III. Design Issues 230
IV. Implementation Issues 234
V. Concluding Section 237

Summary of the Roundtable Discussion

I. Introduction 239
II. Summary of Proceedings 239
III. Conclusions and Key Issues/Areas of Concern 254

Commentary on Evaluation of School-to-Work
by Stephen F. Hamilton 257

Epilogue: Synthesis of the Papers and Discussion

I. Introduction.. 263
II. The Assumptions Underlying School-to-Work 263
III. Design Issues 268
IV. Implementation 275
V. Conclusions 277

Bibliography 281

Preface

This volume brings together the reports and discussion connected with a roundtable convened by the Departments of Education and Labor and the National School-to-Work Office to discuss issues surrounding the conduct of a net impact evaluation of school-to-work. The volume is organized in roughly chronological order. The introduction summarizes the school-to-work approach and the intentions of the School-to-Work Opportunities Act and describes evaluation activities which have been commissioned separately from a net impact effort. Following that, in alphabetical order, are six papers commissioned to provide springboards for the roundtable discussion. These papers were distributed to all roundtable participants before the roundtable met.[1] The. roundtable met on February 23, 1996 in Washington, D. C. A summary of the discussion follows the six commissioned papers. Comments submitted by Steve Hamilton, who was invited to the roundtable but was unable to attend, follow the summary of the discussion. The volume concludes with a synthesis of the papers and discussion.

The participants in the roundtable are listed following this Preface. Eileen Pederson of the Department of Labor had overall responsibility for organizing the roundtable. Edwin Ridgway and Gloria Williams of KRA Corporation handled the logistical arrangements. Nevzer Stacey of the National School-to-Work Office led the discussion and William Goodwin of the Department of Labor was the discussion facilitator. Lloyd Feldman, a consultant to KRA, prepared the summary of the discussion. Daniel Ryan of the Department of Labor prepared the introduction and epilogue for this volume. David Goodwin of the Department of Education provided helpful comments on this volume as well as contributing to the planning of the roundtable.

---

[1] Authors were allowed to revise their papers after the roundtable. Thus the papers contained herein differ, in some cases substantially, from those distributed prior to the roundtable.

List of Participants

Gary Burtless
Brookings Institution

Charles Dayton
Foothill Associates

Robert Glover
Univ. of Texas at Austin

William Goodwin
U.S. Dept. of Labor

David Goodwin
U.S. Dept. of Education

Karen Greene
U.S. Dept. of Labor

James Heckman
Univ. of Chicago

Robinson Hollister
Swarthmore College

J.D. Hoye
National School-to-Work Office

Martha Huleatte
New Jersey Dept. of Education

Christopher King
Univ. of Texas at Austin

David Lah
U. S. Dept. of Labor

Rebecca Maynard
Univ. of Pennsylvania

Robert Moffitt
Johns Hopkins Univ.

Eileen Pederson
U.S. Dept. of Labor

Marion Pines
Johns Hopkins Univ.

Hillard Pouncy
Univ. of Pennsylvania

Peter Rossi
Evaluation Design and Analysis

Daniel Ryan
U.S. Dept. of Labor

Nevzer Stacey
National School-to-Work Office

Reid Strieby
Bronx Community College

Ricky Takai
U.S. Dept. of Education

Raymond Uhalde
U.S. Dept. of Labor

Doris Werwie
U.S. Dept. of Education

Joseph Wire
Office of Management and Budget

Introduction: School-to-Work And School-to-Work Evaluation and Research

I. Introduction

The School-to-Work Opportunities Act of 1994 requires the Secretaries of Education and Labor to conduct an evaluation of school-to-work. The evaluation is to "track and assess the progress of implementation of State and local programs and their effectiveness." In implementing this provision, the Secretaries have commissioned a process evaluation which will track and assess implementation progress. However, assessing the effectiveness of school-to-work raises many challenging issues which cannot be addressed by a process evaluation. Specifically, to measure the effectiveness of school-to-work in terms of learning gains, high school graduation rates, and earnings, as called for in the Act, requires a net impact evaluation -- measurement of these outcomes against what they would have been had the School-to-Work Opportunities Act never been implemented. The challenges of designing and conducting such an evaluation are addressed in this volume.

As background for the remainder of the volume, this introductory chapter first summarizes the intent and provisions of the School-to-Work Opportunities Act. It then summarizes the evaluation and other data collection activities related to school-to-work which are already underway. The final section offers conclusions about the gap in current evaluation activities -- the need for a net impact evaluation.

II. The School-to-Work Opportunities Act of 1994

The School-to-Work Opportunities Act of 1994 provides a national framework to be implemented by all States that will enable youth to "identify and navigate paths to productive and progressively more rewarding roles in the workplace." In enacting the legislation, Congress listed ten issues the Act is intended to address:

- Three-fourths of high school students in the United States enter the workforce without the academic and entry-level occupational skills necessary to succeed.

- A substantial number of youths, especially disadvantaged students, students of diverse racial, ethnic, and cultural backgrounds, and students with disabilities, do not complete high school.

- Unemployment among youths in the United States is intolerably high, and earnings of high school graduates have been falling relative to earnings of individuals with more education.

- The workplace is changing in response to heightened international competition and new technologies. Such forces are shrinking the demand for and undermining the earning power of unskilled labor.

- The United States lacks a comprehensive and coherent system to help its youths acquire the knowledge, skills, abilities, and information about and access to the labor market necessary to make an effective transition from school to career-oriented work or to further education and training.

- Students can achieve high academic and occupational standards and many learn better and retain more when the students learn in context, rather than in the abstract.

- While many students in the United States have part-time jobs, there is infrequent linkage between such jobs and the career planning or exploration, or the school-based learning, of such students.

- The work-based learning approach, combined with school-based learning, can be very effective in engaging student interest, enhancing skill acquisition, developing positive work attitudes, and preparing youths for high-skill, high-wage careers.

- Federal resources currently fund a series of categorical, work-related education and training programs that are not administered as a coherent whole.

- In 1992 approximately 3,400,000 youths in the United States ages sixteen through twenty-four had not completed high school and were not currently enrolled in school; a number which indicates that these young persons are particularly unprepared for the demands of a twenty-first century workforce.

In summary, the School-to-Work Opportunities Act is intended to reform the nation's education and training system so that students are better prepared to enter today's workforce and compete for high-skilled jobs.

The authors of the Act stressed that school-to-work is to be a systemic reform. For example, the Senate report on its version of the School-to-Work Act stated:

> "School-to-work programs represent a fundamentally different approach to teaching and learning that links the school and community. Such programs must therefore be considered an integral part of K-12 education reform. The committee believes that we cannot change education in a piecemeal fashion. All aspects of the system -- standards and assessments, curriculum, teacher development, and governance -- must be addressed. These three legislative efforts [School-to-Work Opportunities Act, Goals 2000, Elementary and Secondary Education Act reauthorization] are based on the premise that a comprehensive approach to reform is needed."

The report also noted that the Act's success depends on employers, as well as schools, taking on new roles:

> "Meeting this challenge will require significant changes in the relationship between schools and employers and in the restructuring of academic and vocational learning. For example, many businesses are reluctant to invest in the long-term training of young people and many school officials are unwilling to share academic decision-making with private employers."

To address these issues, the statute identifies five program requirements which may be treated as the defining elements of a school-to-work transition system:

- integration of academic and vocational learning combining both school- and work-based learning and effective linkages between secondary and post-secondary schooling,
- defined career majors,
- incorporation ofschool-based learning, work-based learning and activities connecting the two,
- provision to students of experience in all aspects of the industry the students are preparing to enter, and
- provision to all students of equal access to the full range of school-to-work program components.

Integration of work-based and school-based learning and linkages between secondary and post-secondary education addresses the belief that there are inadequate and infrequent linkages between jobs held by students and the students' future careers. It is believed that such linkages will make school-based learning more relevant and thereby motivate students to learn. It is also believed these linkages will enable youths to avoid the "churning" which characterizes the current school-to-work transition .

Career majors are defined as "a coherent sequence of courses or field of study that prepares a student for a first job." The statute further stipulates that career majors typically include at least two years of secondary education and one or two years of post-secondary education and result in the award of certifications of skills attainments, including diplomas. Students are to make an initial selection of a career major by the beginning of the eleventh grade. A program of career exploration and counseling is to be provided beginning not later than the seventh grade to prepare students to make this decision.

The three components of every school-to-work program are to be: school-based learning, work-based learning and connecting activities. The school-based learning component is centered on the student's career major and intended to meet "the same challenging academic standards established for all students in their state." Further, according to the Conference report on the bill, "participating students who complete a school-to-work program of study at the secondary level should be adequately prepared to enter an associate or baccalaureate program as well as to earn a skill certificate."

The Senate report identifies the aims of work-based learning as "to give practical meaning to academic concepts and to transform traditional instruction into learning experiences." It goes on to say, "Work-based learning places students in actual work settings where students learn real, functional and sustainable skills." The work-based learning component must include work experience, workplace mentoring, and instruction in "general workplace competencies."

Connecting activities are intended to connect the school- and work-based learning components. Such activities include: matching students with work-based learning positions, providing technical assistance to employers in designing work-based learning, and linking school-to-work activities with employer and industry strategies for upgrading skills. The Act's description of connecting activities suggests the important role employers play in school-to-work. As the Dayton paper in this volume points out, "Unless employers play the needed role, the intended partnerships between education and business will be only empty shells, unable to fill the roles necessary to accomplish the job."

In requiring school-to-work to provide training in all aspects of an industry, the Senate noted, "The more aspects of an industry a student is exposed to and is educated in, the broader his or her understanding of that industry will be and the better the educational experience. Broad instruction in all aspects of an industry also broadens students' career options by exposing them to issues that cut across occupations and industries." Dayton points out that "given the changing nature of the workplace, at the secondary level students need to understand the broad features of an industry rather than be trained for specific jobs within it. They need to learn skills that will be transferable from one company to another."

The final program requirement is that all students be provided equal access to all school-to-work activities. The Act specifically mentions disadvantaged students, students with diverse racial, ethnic or cultural backgrounds, students with disabilities, students with limited English proficiency, migrant children, school dropouts, and academically talented students. Specifically, in urging the legislation, "advocates focussed on college-bound youth and the possibility that the program might blur the boundaries between college-bound and career-track programs" (Pouncy and Hollister).

The governance structure for school-to-work foresees four levels of responsibility. At the Federal level, general guidance, technical assistance and funding are to be provided by a joint effort of the Departments of Education and Labor. The National School-to-Work Office, comprised of staff from both Departments, has been formed to carry out these functions. The Federal role also includes conducting research and demonstration programs, establishing a system of performance measures in collaboration with the states, conducting a national evaluation of school-to-work, and collecting and disseminating information on successful school-to-work approaches. The Secretaries of Education and Labor are required to submit an annual report to Congress which includes evaluation findings. The Federal role in school-to-work will effectively end with fiscal year 1999 -- the last year for which funds can be appropriated under the Act.

State governments are to establish local partnerships, organize the participation of the various entities involved in a school-to-work system and coordinate and support local planning and development activities. Dayton points out that, "A central problem in many states and localities in the past has been the lack of collaboration among agencies and programs concerned with preparing youth for work." The School-to-Work Opportunities Act addresses this issue by giving responsibility for local school-to-work programs to "partnerships." These partnerships are to include employers and representatives of local education agencies, labor and students. A variety of other entities may also be included in a partnership. Finally, the "front line" in school-to-work -- the point where the students encounter school-to-work -- consists of local schools collaborating with employers and others.

The implementation of school-to-work systems is to begin through the distribution of "seed money" by the Federal government. All States plus Puerto Rico and the District of Columbia were given "development grants" ranging from $200,000 to $750,000 in January 1994 to begin planning their school-to-work efforts. In June 1994, eight States were chosen competitively to receive "implementation grants."[2] An additional 19 States were awarded implementation grants in September of 1995. Further, approximately 100 localities have received local partnership grants directly from the Federal government and Urban/Rural Opportunity Grants targeted to high poverty areas; roughly half of these grants are in states that have not received implementation grants. As of this writing, a grant competition is being held to award implementation grants to approximately ten additional states. Once states have received implementation grants, they may receive additional funds for up to five years to complete implementation.

III. School-to-Work Evaluation

Among the responsibilities of the Secretaries of Labor and Education is to evaluate the implementation of school-to-work in terms of effectiveness in:

- developing and implementing State programs,
- participation in school-to-work by employers, schools, students and school dropouts,
- addressing the needs of all students and school dropouts,
- improving achievement of participating students in terms of academic learning gains; attainment of high school diplomas, skill certificates, and postsecondary degrees; attainment of experience in and understanding of all aspects of an industry; placement and retention in further education or training; and job placement, retention and earnings,
- meeting state goals to ensure opportunities for young women to participate in nontraditional employment, and
- meeting the needs of employers.

---

[2] See Reisner et al. for a discussion of these states early experience with school-to-work.

To address this mandate, the Departments of Education and Labor and the National School-to-Work Office have developed an evaluation strategy to answer three questions:

- What progress have states and local communities made in establishing school-to-work systems that fundamentally change how young people are educated and prepared for careers?
- Are trends in participation, customer satisfaction and available resources indicative of school-to-work's long-term viability?
- What are the educational and labor market outcomes associated with school-to-work systems?

The major effort currently underway to address these questions is a comprehensive evaluation of school-to-work system implementation. The specific objectives of this evaluation, which began in September 1995, are to:

- Document the progress made at the state and local levels in implementing school-to-work systems.
- Identify promising practices and barriers to implementation progress.
- Describe the participation of students, schools, employers and other organizations in school-to-work partnerships and activities.
- Describe students' educational and employment outcomes and how they are changing as school-to-work systems evolve.

To pursue these objectives, the evaluation includes three major components:

- A survey of local partnerships will be conducted in the Fall of 1996, 1997 and 1999. The survey will cover partnerships awarded grants by the 27 states with implementation grants as well as partnerships awarded direct Federal grants. It will collect information on partnership organization, school-to-work program features, links between secondary and post-secondary education, employer participation, and aggregate measures of student participation in particular program activities.
- Detailed case studies of program implementation and factors affecting program design and progress will be conducted in 1996, 1997, and 1999. Data for these case studies will be

collected through site visits to four local partnerships in each of eight states with implementation grants and 10 local partnerships with direct Federal grants.

- In the eight states in which case studies are conducted, surveys of three cohorts of twelfth graders, selected and interviewed in the Spring of 1996, 1998 and 2000. A follow-up interview will be conducted with each cohort eighteen months after the end of their senior year. Questions will cover the students' high school experiences and, at follow-up, post-secondary education and labor market experiences. High school transcripts will be collected to provide further information about their course-taking patterns and academicperformance.

In a separate effort to track broad national trends, questions related to school-to-work experience have been added to the National Longitudinal Survey of Youths (NLSY). Beginning in 1996, the NLSY will annually survey a national sample of youths ages twelve through seventeen. In addition, a survey of school administrators at 8,500 schools which have sample members in the NLSY will collect data on school policies and practices related to school-to-work activities. The school administrators' survey will be conducted every other year beginning in 1997.

To further measure implementation progress, the states in collaboration with the Federalgovernment are to implement a system for measuring performance. The progress measures activity will develop a core set of performance measures that provide basic information on program implementation and establish a system of assessment. Partnerships are asked to identify on a "progress measures chart" the level of involvement in school-to-work at elementary, middle and secondary schools. This system will, at a minimum, provide standardized definitions of the most important school-to-work terms and, thereby, contribute to both the school-to-work dialogue and evaluation.

Finally, a survey of a nationally representative sample of employers was conducted in 1994 and 1996 and will be repeated in Spring 1998 and October 2000. It will gather information on the extent of employer involvement with schools and school-to-work and measure the benefits to employers of such involvement.

IV. Conclusion

In this concluding section, we highlight those aspects of school-to-work which the papers and discussion point to as the most important contextual factors for a net impact evaluation of school-to-work. We then discuss the contributions of the already implemented evaluations and research as they relate to a net impact evaluation.

The overarching contextual feature of school-to-work is the administrative structure which Pouncy and Hollister describe as an "administrative sandwich." The devolvement of responsibility for design of school-to-work systems has resulted in a "let a thousand flowers bloom" approach to school-to-work implementation. Specifically, as Glover and King point out:

> "That school-to-work models and treatments feature enormous within- and between-intervention variation, with many unknown hybrids of the programs operating in practice, makes conducting net impact evaluations of such efforts quite difficult. These problems are further exacerbated by lack of consistent Federal component definitions and by the fact that State and local partnerships largely occupy the driver's seat for implementing these programs, resulting in considerable inter-area variation as well."

As will be seen throughout this volume, this has implications for both the issues a net impact evaluation should address and the evaluation design. For example, Burtless sees this as an opportunity for the evaluation:

> "The variety of program models that states and localities have developed opens rich possibilities for the national evaluation. Investigators can attempt to determine whether one program model has been consistently more successful than others. They can try to estimate the relative effectiveness of small variations in treatment within the same basic model."

He further notes that,

> "Since many State and local programs are now in their infancy, program administrators are eager to learn about the relative effectiveness of different approaches. For that reason, there is likely to be large payoff from research that reliably distinguishes between successful and unsuccessful approaches as well as from research that shows whether certain kinds of students and dropouts can benefit from particular school-to-work models."

Burtless also points out a "second and equally important" contextual feature of the School-to-Work Opportunities Act: "[T]he Act envisioned a fundamental change in the nation's school-to-work system." Burtless mentions that such fundamental change "alters the expectations of all major actors in the system." Such change will take many years -- perhaps decades -- to be fully realized. However, it is doubtful that information from an evaluation will still be useful after that long a time. Once such change is fully complete, school-to-work will be too institutionally entrenched to benefit from the information a net impact evaluation would provide. Thus, there is a trade-off between the scope and timeliness of an evaluation.

Dayton separates the policy issues raised by the School-to-Work Opportunities Act into ten categories.

1. Collaboration -- the need for many stakeholders to work together.
2. Fostering systemic change in schools -- fundamental changes in how schools operate.
3. Establishing work- and community-based learning -- the necessity of expanding the settings for school-to-work preparation beyond school campuses.
4. Professional development, resources and technical assistance -- the implications of these broad ranging changes for needed support.
5. Standards, assessment and credentialing -- their role in defining and driving the school-to-work movement.
6. Technology and the high performance workplace -- how changes in the workplace must be reflected in the classroom.
7. Teaching all aspects of the industry -- keeping the career focus broad, not job specific.
8. Pedagogy -- whether changes in instructional methods lead to improved student outcomes.
9. Serving all students -- the advantages of school-to-work for fairly serving every type of student.
10. Employer involvement -- the critical role of employers in the school-to-work effort.

Dayton points out that these policies are "far reaching, covering a host of educational reforms." Yet a comprehensive evaluation of school-to-work would address all these issues.

While all of Dayton's categories could be subjects for net impact evaluation, the first five categories primarily raise implementation issues which are adequately addressed in the evaluation already underway. Categories 6 through 8 outline changes in instructional methods and subject matter.

While the evaluations described above can tell us if such changes are taking place, they cannot tell us if such changes lead to improved student outcomes. Similarly, the implementation evaluation and student surveys can tell us if all types of students are being served by school-to-work -- category 9; they cannot tell us if all students are benefitting from school-to-work. Finally, category 10 raises the key roles employers must play if school-to-work is to be a success. The employer survey mentioned above and the implementation evaluation will examine the extent to which employers are playing these roles. They leave open the question of what benefits are obtained by employers, schools and students when employers take on these roles.

Burtless recommends that a "helpful" employer study would address two questions:

- What do participating firms gain from their participation and what changes in school-to-work would make continued participation by employers desirable to the employer and make school-to-work effective in meeting employer needs?
- Why do non-participating firms fail to participate in school-to-work?

Both of these questions could be answered by an appropriately designed survey of employers rather than a net impact study. However, a study of school-to-work's net impact on employers -- if it found positive impacts -- would be a powerful tool for selling employers on participation in school-to-work.

Several authors categorize evaluations by type. Dayton, for example, describes three types of evaluation which could be used to assess effectiveness in addressing these categories: implementation evaluation, policy evaluation and evaluation of participant effects. The evaluation already underway meets most of the aims Dayton describes for implementation and policy evaluation. Thus, the net impact evaluation should concentrate on categories 6 through 10 --'school-to-work's net impacts on students and employers.

Evaluating Early Program Experiences in the School-to-Work Opportunities Act: Policy and Design Issues

Gary Burtless
Senior Fellow, Economic Studies
The Brookings Institution

I. Introduction and Summary

The School-to-Work Opportunities Act of 1994 (STWOA) was intended to improve the transition of young Americans from school into the workplace. The Act's overall goal is to encourage States and localities to organize school-to-work systems that link learning in school with learning in the workplace.[1] States and localities are expected to forge links among schools, local employers, and postsecondary education and training institutions in an attempt to improve secondary students' and dropouts' access to training opportunities among local employers and postsecondary institutions. Under awell-functioning school-to-workprogram, students should benefit through improvedlearning in school, better preparation for postschool employment, and superior access to postsecondary education and training opportunities. Employers should gain from a better trained workforce in general and from improved information about the preparation and skills of individual young people they have helped to train.

The purpose of this paper is to consider formal assessment of the success of programs that are funded under the STWOA. The paper will describe the overall objectives of a sensible evaluation, several important subsidiary issues that a sound assessment would consider, and two basic approaches toward evaluation of the effectiveness of STWOA-financed programs. To meet the goals of the evaluation that Congress described in the Act, two basic approaches will be needed. First, several congressionally mandated evaluation goals can be attained through intelligent use of the performance measures that States and localities are required to report under the Act. A well-functioning management information system will provide the Secretaries of Education and Labor with timely information about the number, demographic composition, and academic and training progress of students and dropouts who participate in STWOA-financed programs. This information can be used to assess the effectiveness of State and local programs in meeting several goals of the Act.

---

[1] For a good discussion of the STWOA, as well as background information about previous school-to-work programs, see Office of Technology Assessment, 1995.

Second, two of the congressional goals described in the Act can be assessed only by using formal statistical evaluation procedures. Federal public servants or contractors must develop experimental or quasi-experimental data sources and statistical procedures to determine whether State and local programs have been effective in helping improve student outcomes. This undertaking is very difficult for two reasons.

One problem is that the programs financed under the STWOA are extremely decentralized. They are not administered by a single public authority. Rather, they are separately administered by State authorities and, in most cases, also by numerous local authorities. Therefore, evaluators will find it extremely costly and perhaps impossible to evaluate the overall effectiveness of STWOA programs for a representative sample of program participants. Instead, the formal statistical analysis of program success must be restricted to a sample of participants drawn from a sample of the State and local authorities that run programs financed under the Act. Because each State and most local authorities are free to develop unique and uncoordinated programs to meet the goals of the Act, the people who design and administer the national evaluation will have little scope to vary the programs in a systematic way so that the most important policy questions can be addressed in the study.

A second and equally important problem is that framers of the Act envisaged a fundamental change in the Nation's school-to-work system. If fundamental change actually occurs, it will alter the expectations of all major actors in the system--students, dropouts, educators, and employers. When these expectations change, the behavior of the main actors will also change, affecting the Nation's school-to-work system and the institutions that participate in it. In the long run, actors' behavior will differ from what we observe when the STWOA programs are initially implemented.

Unfortunately, the expectational changes will occur over a lengthy period, probably one that spans at least one or two decades. However, an evaluation of the STWOA must be completed by

September 30, 1998, under Section 401(b) of the Act. No one seriously believes that the expectations or behaviors of the major actors will stabilize by September 1998, much less by 1997 or early 1998 when program experiences under the STWOA will be measured. Therefore, it will be impossible to ascertain the full response to programs under the STWOA in a reliable way within the evaluation deadline specified by the Act.

However, evaluators will have an opportunity to measure partial responses to the programs during the next two or three years. In many cases, obtaining a reliable measure of initial program effectiveness will be useful in indicating which approaches to school-to-work programs are most promising. For that reason, an early or partial evaluation of STWOA programs can have great value to Federal, State, and local policymakers.

II. Conceptual and Policy Issues

Section 402 of Public Law **103-239—the** School-to-Work Opportunities Act-requires the Secretaries of Education and Labor to (1) collect information on performance outcomes under the STWOA and (2) complete a national evaluation of programs funded under the Act no later than September 30, 1998. Congress was specific in defining the nature of the evaluation. The Secretaries were obligated to "track and assess the progress of implementation of State and local programs and their effectiveness based on measures such as those . . . described in subsection (a) [of Section **402]"** (italic supplied). This language implies that two kinds of study will be needed-an implementation analysis describing State and local success in establishing and administering school-to-work programs and an evaluation study showing whether the programs were effective in achieving the goals of the Act. This paper focuses on the latter obligation, which will be more difficult to meet. Completion of the implementation analysis, because it is mainly descriptive, should be straightforward.

1. Congressional Goals for the National Evaluation

Congress provided a guide to its evaluation intentions in subsection (a) of Section 402 of the Act, in which it describes the kinds of performance measures the Secretaries of Education and Labor must obtain. The performance measures are divided into the following six categories:

1. Progress in developing and implementing State school-to-work plans as well as specific program components described elsewhere in the Act

2. Participation in programs funded under the STWOA on the part of employers, schools, students, and dropouts, including information about specific kinds of groups of students (i.e., classified by race, ethnicity, socioeconomic background, English language proficiency, disability status, and level of academic talent)

3. Progress in establishing and implementing strategies that address the needs of students

4. Progress in ensuring that young women participate in STWOA-funded programs, including programs that prepare young women for employment in nontraditional occupations (i.e., occupations traditionally held by men)

5. Outcomes for students and school dropouts who participate in STWOA-funded programs, classified by demographic group, including outcomes such as the following:

   - Academic learning gains

   - School retention, school graduation, attainment of a skill certificate, and attainment of a postsecondary degree

   - Experience in and understanding of industries and occupations in which students obtain training

   - Placement and retention in higher levels of education and training, especially in the area in which the student received STWOA-financed training

   - Job placement, retention, and earnings, especially in the area in which the student received STWOA-financed training

6. Success in meeting the needs of employers

The national evaluation described in subsection 402(b) of the Act should be clearly integrated with the design and collection of performance measures described in subsection 402(a). However, because the evaluation requirement requires the Secretaries of Education and Labor to assess the effectiveness of State and local programs, the national evaluation must go beyond the simple task of tabulating the performance measures required under subsection 402(a).

Simple Benchmarks of Program Effectiveness

Effectiveness in attaining some of the goals of the legislation can be assessed by comparing State and local accomplishments under the STWOA with a realistic benchmark that defines a plausible level of achievement. For example, participation by young women in STWOA-financed programs can be compared with participation by young men. If young women and young men have similar rates of participation, the Secretaries and Congress could conclude that at least one goal of the STWOA has been achieved. In addition, if young women enroll in education and training courses for historically male occupations in about the same percentage as young men, the Secretaries and Congress could conclude that another important objective has been achieved.

It is unlikely, of course, that participation rates in STWOA-funded programs will be exactly the same in every demographic and academic achievement group described in subsection 402(a)(2). When participation rates differ, it is important that authors of the evaluation report attempt to explain the differences. Nonetheless, in many cases, Congress has implicitly provided a natural and simple benchmark for assessing whether State and local programs have been effective in achieving the STWOA's broad public purpose. In these cases, the national evaluation of State and local program performance can be based on program performance measures collected under subsection 402(a) of the Act.

More Complicated Benchmarks of Program Effectiveness

In other cases, Congress clearly has a more complicated kind of assessment in mind. Subsection 402(a)(5) of the STWOA asks for measures of participants' academic learning gains, for example. If one goal of the Act is to boost learning gains, evaluators must establish whether STWOA-funded programs had a beneficial effect on student learning. Reliable measurement of the effect requires some basis for comparison. In some evaluations, a before-and-after comparison provides acceptable results. The Department of Transportation tests the structural integrity of automobile interiors by driving sample cars into a fixed barrier and then ascertaining the damage to a crash-test dummy. The implicit benchmark for comparison is the undamaged state of the dummy before the crash occurred. However, a before-and-after comparison is not always appropriate or feasible.

In the case of academic learning, we can be sure that most STWOA participants would enjoy gains on standardized tests in secondary school, whether they participated in STWOA-funded programs. Therefore, the evaluator must try to answer a different kind of question than the one posed by the Department of Transportation in its automobile crash tests. Instead of posing the question, "How does the student's performance on an academic test change over the course of enrollment in a school-to-STWOA-funded program?" the evaluator must ask "How does the change in the student's academic performance compare with what the change in performance would have been if the student did not participate in the STWOA program?"

The problem, of course, is establishing a credible basis of comparison. Each student must make a choice to participate or to decline participation in a STWOA program. When the student graduates or leaves high school, the evaluator can measure the student's academic performance as a participant or as a nonparticipant in the STWOA, but not as both. It is therefore not obvious what benchmark for comparison can be used to assess the academic learning gains of students who participate in the STWOA-financed programs. In essence, this is the evaluation problem faced by the Federal government in measuring the effectiveness of programs financed under the STWOA: Investigators cannot observe what a participant's performance would have been in the absence of the program.

2. Research Hypotheses

This discussion suggests the national evaluation must use two basic approaches toward defining an appropriate benchmark for assessing program effectiveness. Different goals of the Act should be evaluated with this distinction in mind. For one set of goals, it is appropriate to set a benchmark by considering evidence that is comparatively easy to observe (program participation rates among different demographic groups or among different categories of students defined by their level of English-language proficiency or scholastic aptitude). For the other set of goals, evaluators must use sophisticated statistical techniques to ascertain an appropriate benchmark. For example, to determine whether a STWOA-funded program has been effective in boosting academic performance, evaluators must first establish how well students who enrolled in the program would have done if the program had not been established. This is a benchmark level of performance that cannot be directly observed and therefore must be ascertained through somewhat involved statistical study.

It is useful to describe the research hypotheses that should be examined in the evaluation and to divide them into the two categories suggested by the previous discussion.

Research Questions With a Straightforward Performance Benchmark

1. Has a State (or locality) developed a school-to-work plan that meets the requirements of the STWOA? Has it implemented the plan?

The performance benchmark in this case is straightforward. For the first part of the question, the benchmark is defined by the Act itself, which describes the conditions that State plans must meet to conform with the STWOA. Presumably, some States have developed plans that meet all of the tests implied by the legislation; others have failed to develop plans or have developed plans that fall short in some crucial areas.

Implementation of the plans can be examined using the performance information that States are required to provide to the Secretaries of Education and Labor. A simple-but unrealistic-standard would define an effective State as **one** where 100 percent of the school systems in the State have school-to-work programs in place that conform to the State plan. A more realistic benchmark would be the actual performance of the State that ranks number 10 in implementation. Nine States have a higher percentage of school systems offering a school-to-work program that conforms to the State plan, whereas 40 States have a smaller percentage of school systems with a conforming program.

2. Has a State (or locality) succeeded in achieving a high level and equitable distribution of participation in school-to-work programs?

In this case, defining a benchmark for measuring effectiveness is somewhat less obvious. Is a 30-percent rate of student participation too low? Or is it too high? In the initial stages of implementing the STWOA, it is probably safe to say that a high rate of participation indicates better program performance than a low rate. It may make sense to ascertain the participation rate among eligible eleventh and twelfth graders in all the States and then define effective performance as the level of participation achieved by the State ranking tenth in the cross-State distribution.

The benchmark for assessing whether the distribution of participation is equitable is easier to establish. As mentioned earlier, an equitable distribution might be one in which all groups in a State's (or locality's) student and dropout populations have similar participation rates.

3. Has a State (or locality) succeeded in implementing strategies that ensure equitable participation by young women in STWOA-funded programs, especially in training for nontraditional occupations?

The benchmark for determining whether States and localities have implemented successful programs is the participation rate of young men in school-to-work programs. If young women achieve similar rates of participation, especially in training for male-dominated occupations, the goal implied by the legislation has been met-unless, of course, both men and women have negligible participation rates. Evaluators should also ensure that the training received by young women has been as effective in improving learning and employment for young women as it has been for young men (see below).

Research Questions Requiring a Complex Performance Benchmark

4. Has a program improved academic performance in comparison with performance under the traditional system? How do the performance effects differ across demographic and other kinds of groups?

5. How has the program affected school retention rates? School completion rates? Attainment of skill certifications? Attainment of postsecondary degrees? Do the effects differ across demographic and other kinds of groups?

6. How has the program affected participants' knowledge of the skills needed to work in a variety of industries and occupations? Have increases in job-specific knowledge helped bring participants' skills up to the standard for entry-level positions? Up to the standard for more demanding positions? Do the effects differ across demographic and other kinds of groups?

7. Has the program improved participants' rate of job entry? Rate of job retention? Postprogram earnings? Are the effects, if any, concentrated in jobs that require the skills taught in the school-to-work program? Do the effects differ across demographic and other kinds of groups?

8. Has the program met the needs of students? Of employers?

Most of these research questions demand a longer period of study than that provided by the September 1998 evaluation deadline in the STWOA. Students who enter a new school-to-work program in fall 1996 may not even have completed high school by September 1998, so it is unrealistic to think the evaluation can provide reliable estimates of the effect of the program on school completion rates or postsecondary enrollment rates. The national evaluation required by the STWOA should represent the first installment of a longer term research project to monitor the success of school-to-work programs. The remainder of this report assumes that policymakers ultimately hope to obtain measures of the long-term educational and employment impacts of school-to-work programs for students who are entering these programs in the next few years. This approach will require data collection and analysis for several years after the September 1998 evaluation report has been completed.

Additional Research Issues

In addition to the major research questions, which are mentioned or implied in the Act itself, evaluators should attempt to learn whether certain school-to-work program models yield better results than others. In the long run, this kind of information will have much greater value to policymakers than the answers to questions 1 through 8, above. An evaluation report that is completed by September 1998 cannot offer much guidance about the long-term benefits of a reformed school-to-work system. Students, employers, and school administrators will have too little experience under the new programs for policymakers to draw reliable conclusions about the longer term effects of mature programs that are run by experienced educators and employers. However, careful analysis of results from a variety of good programs might reveal whether certain program models show special promise.

The Office of Technology Assessment has identified the following six broad models of workplace-centered learning: youth apprenticeship, clinical training, cooperative education, school-to-apprenticeship training, school-based enterprises, and career academies.[2] Within each model, schools and employers may take different approaches to training and rewarding participants for the work they do. Some programs emphasize the involvement of adult mentors for each trainee. Others emphasize classroom instruction in schools. Still others may require students to spend a large proportion of the academic year-perhaps fifteen or more hours a week-in a job. Some programs require trainees to be paid for workplace-based learning that occurs on a job, whereas other programs offer no pay to trainees who work. Experience has shown that one of the toughest challenges facing workplace-centered learning is recruitment and retention of local employers to participate in school-to-work programs. Some States or localities may develop innovative methods to attract and keep employer participation.

The variety of program models that States and localities have developed opens rich possibilities for the national evaluation. Investigators can attempt to determine whether one program model has been consistently more successful than others. They can also attempt to estimate the relative effectiveness of small variations in treatment within the same basic model. Currently, as the Office of Technology Assessment shows, analysts and policymakers have little reliable information to show whether one approach to workplace-based education is better than another. Because many State and local programs are now in their infancy, program administrators are eager to learn about the relative effectiveness of different approaches. For that reason, there is likely to be a large payoff from research that reliably distinguishes between successful and unsuccessful approaches, as well as from research that shows whether certain kinds of students and dropouts can benefit from particular school-to-work models. An important focus of the national evaluation should be measurement of the relative effectiveness of different program models.

---

[2] Office of Technology Assessment, op cit., pp. 58-59. For a discussion of experience under the youth apprenticeship and school-to-apprenticeship models, see Thomas R. Bailey and Donna Merritt, 1993.

Evaluators find it difficult to measure the relative effectiveness of different program models for two main reasons. First, the evaluation problem makes it difficult to measure the effect of any treatment that is offered outside of a controlled experiment (see below). However, even if the evaluation problem were solved, the investigators' problems would not be over. Analysts are rarely given the opportunity to observe a clean and simple test of the relative effectiveness of two program models. Community A may establish career academies, and community B may offer a modified version of cooperative education. The relative success of the two kinds of programs in the two communities does not offer a reliable guide to the relative success of the two approaches if they were adopted in other communities. Suppose participant earnings rose 15 percent in community A but only 5 percent in community B. It does not follow that career academies are generally more effective than cooperative education programs. Perhaps community A is blessed with outstanding school administrators. If these administrators had been given the task of organizing and managing a cooperative education program, they might have created a program that boosted participants' earnings by 15 percent-exactly the same earnings gain that was achieved in community A's career academies. Alternatively, the demographic characteristics of program participants may have differed in the two communities, making it hazardous to compare earnings gains in the two programs.

Despite the difficulties of determining relative program effectiveness, the lessons learned on this subject could be the most important ones we learn from the national STWOA evaluation. Given resource constraints and estimationdifficulties, it is impossible to measure the relative effectiveness of every interesting model or of every interesting variant of a basic school-to-work model. Nonetheless, the evaluation would be greatly improved if the Secretaries of Education and Labor selected two or three important program variations and devoted a portion of the evaluation budget to measuring the relative effectiveness of these variants. Good research design will be needed to ensure reliable results. In selecting the program variants to be examined, the Secretaries should rely on the judgments of officials and researchers who are most familiar with the practical design and administrative problems faced in school-to-work programs.

III. Design Issues

Evaluators should have little difficulty answering research questions 1 through 3. The estimation problems in answering questions 4 through 8 and the questions about relative program effectiveness are more formidable. In measuring the effectiveness of programs financed under the STWOA, investigators face a crucial problem: They do not observe what participants' educational performance or employment or earnings would have been in the absence of the program.

1. Quasi-Experimental and Classical Experimental Approaches to Evaluation

To solve the evaluation problem, good analysts have used two basic strategies-quasi-experimental comparisons and classical experiments.[3]

Quasi-Experimental Methods

In the first approach, an investigator compares the average academic performance gain of STWOA participants with that of a hand-picked group of students who, for some reason, failed to participate in a STWOA-financed program. Naturally, the investigator will want this comparison group to be very similar to the group of students and dropouts who enrolled in STWOA-funded programs. If the investigator cannot plausibly argue that the two groups are identical, he or she must make a statistical adjustment to remove all outcome differences between the two groups, except the difference in outcome that is caused by participation in the STWOA program.

The principal advantage of quasi-experimental comparisons is that an estimate of program effectiveness can usually be obtained without much interference in the actual operation of the program in question. A simple example can illustrate this point. Suppose community

---

[3] **For basic discussions of the advantages and disadvantages of each approach, see Gary Burtless, "The Case for Randomized Field Trials in Economic and Policy Research," *Journal of Economic Perspectives*, Vol. 9, 2 (spring 1995, pp. 63-84); and James J. Heckman and Jeffrey A. Smith, "Assessing the Case for Social Experiments," *Journal of Economic Perspectives*, Vol. 9, 2 (spring 1995, pp. 85110).**

A institutes a distinctive school-to-work program for its students. Community B, which is similar in most respects to community A, fails to implement a similar program, choosing instead to maintain its traditional academic/general education/vocational schooling tracks.

To estimate the effects of the school-to-work program using quasi-experimental methods, the investigator would compare participants in community A's program with students or dropouts in community B who are similar to participants in community A. (For example, comparison-group students in community B might be chosen to have the same mix of race, gender, English-language proficiency, and eighth-grade test scores as the students in community A who actually participated in the program.) This kind of statistical comparison does not require that the investigator interfere with educational policies in either community. Instead, the evaluation requires only a data collection plan that provides identical information about characteristics and outcomes for students in the two comparison groups and an analysis plan that statistically controls for differences between the two groups.

Investigators face two main problems in implementing the quasi-experimental comparison strategy. First, it is almost always difficult (and often impossible) to find a group of people who are nearly identical to the people who choose to enroll in a program that is aimed at a minority. In the previous example, it might be difficult to identify a community B. Or, if community B is successfully identified, it may be difficult to select students and dropouts in that community who are very similar to program participants in community A. Second, when analysts are forced to use a statistical adjustment procedure to make the two comparison groups appear identical, the results are invariably open to question. No reliable test can assure us that the procedure has actually succeeded in eliminating differences between the two groups, except the outcome difference that is due to program participation.

A common source of bias in statistical evaluation studies is sample selection. Quasi-experimental studies of education programs, for example, usually rely on observations of naturally occurring variation in treatment in order to form estimates of the effects of the program. Analysts typically compare outcomes on standardized tests for participants in the program and for a comparison group of similar people who did not participate in the program. Even if the analysis fully controls for the effects of all measurable characteristics of students in the sample, it is still possible that average outcomes are influenced by systematic differences in the unmeasured characteristics of individuals in the treatment and comparison groups. For example, in the simplest kind of education program, members of the estimation sample are exposed to only two doses of treatment: 1 for people enrolled in the program, and 0 for people who never enrolled. Program participation represents the sample member's (or school administrator's) decision to choose one treatment dose or the other. Obviously, this decision may be affected by unobserved tastes or other characteristics that also affect an individual's later test scores. Because these factors are not known and cannot be estimated, the amount of bias in the quasi-experimental estimate of the program effect will be unknown.

Selection bias is a practical estimation problem in most quasi-experimental evaluation studies. Analysts who use quasi-experimental methods sometimes ignore the problem, implicitly assuming that unmeasured differences between the two groups do not exist or do not matter. Although one or both of these assumptions could be true, the case for believing either of them is ordinarily not very strong. Students who voluntarily enroll in a school-to-work program might be more ambitious than other students, yet a student's ambition is impossible to measure. If personal ambition is correlated with the student's subsequent performance on an academic test, it is unclear what percentage of the average test-score advantage among school-to-work participants is due to the effects of the program and what percentage is due to the higher level of ambition among students who participated in the program. The selection bias could go in the opposite direction, too. Students who are not optimistic about their future success in school may disproportionately enroll in school-to-work programs. If their pessimism is based on a realistic-but unobserved-assessment of their prospects, we should expect that their test scores, if they did not participate in the school-to-work program, would be lower than those of students with identical observable characteristics who chose not to enroll in the program,

Our uncertainty about the presence, direction, and potential size of selection bias makes it hard for competent evaluators to agree on the reliability of estimates based on quasi-experimental methods. If investigators obtain widely differing estimates or if the available estimates are the subject of strong methodological criticism, policymakers will be uncertain about the effectiveness of the program.

Randomized Trials

A second approach to the evaluation problem is to conduct a classical experiment. The critical element that distinguishes classical experiments from other methods of study, including quasi-experimental comparisons, is random assignment of treatments to the subjects involved in the study. A randomized field trial (or social experiment) is simply a classical experiment that takes place outside a laboratory setting, in the environment where social, economic, and educational processes normally occur. In the simplest kind of experiment, a single treatment is assigned to a randomly selected subsample (the treatment group) and withheld from the remainder of the enrolled sample (the control or null-treatment group).

Using the experimental approach, the evaluator would randomly divide students into two groups, one which is eligible to participate in a STWOA-funded program (the treatment group) and another which is ineligible and not permitted to enroll (the control group). To determine whether the program was effective in improving academic performance, the investigator would collect data about both groups and compare the postprogram test scores of students enrolled in the treatment group with postprogram scores of students in the control group. The simple outcome difference between the two groups is a reliable measure of the program's effect if scores have been accurately measured and information from a random sample of both groups has been obtained.

The assignment procedure assures us of the direction of causality between treatment and outcome. Differences in average outcomes between the treatment and control groups must be caused by differences in treatment; differences in average outcomes are not the cause of the observed differences in treatment. Random assignment also removes any systematic correlation between treatment status, on the one hand, and observed and unobserved participant characteristics, on the other. Estimated treatment effects are therefore free from the selection bias that potentially taints estimates based on quasi-experimental comparison-group methods.

Of course, good experiments are neither cheap nor easy to design and run. Experiments cause. three kinds of problems that make them more expensive than typical quasi-experimental studies. They consume a great amount of real resources, especially in comparison with quasi-experimental studies that rely on existing data sources. They may also be costly in terms of time. It is ordinarily more time consuming to design an experiment and to implement policies that ensure the integrity of the treatment and control groups than it is to design a nonexperimental study and identify potential sample members for the treatment and comparison groups.

In addition, and perhaps most important, experiments often involve significant political costs. Unlike quasi-experimental methods, they almost always involve some kind of interference in the normal operations of a school or training institution. Even if school administrators are given unlimited scope to design and implement their school-to-work programs as they wish, they would not be allowed to assign students to the program as they see fit. That choice is left to the investigator or, more accurately, to a device that produces random numbers. School officials are not permitted to enroll control-group members in their school-to-work program, even if they believe these students would enjoy sizable educational benefits from enrollment.

It is obviously more costly to develop, implement, administer, and enforce procedures to guarantee the integrity of random assignment than it is to analyze information about students' performance in a quasi-experimental study. Voters and policymakers are greatly-and properly-concerned about ethical issues raised by experiments. As a result, it is usually easier to persuade policymakers and educators to support quasi-experimental studies than it is to convince them that some students should be systematically denied entry into a potentially beneficial new program for the purpose of obtaining highly reliable statistical estimates.

Experiments Versus Nonexperimental Methods

Classical experiments are often the most cost-effective and reliable way to produce estimates of program effect that are widely accepted among social scientists and believed by policymakers. The most understandable and persuasive experiments usually have simple, straightforward designs. Simple experimental designs have the best chance of producing results that are robust with respect to econometric specification and that are convincing and useful to policymakers. This is their main advantage over quasi-experimental research designs.

The problem with experiments is that they are costly, politically difficult to implement, and time consuming. Moreover, because the Federal government has only limited authority over the school-to-work plans that States and localities may establish, there is little scope for implementing a coherent set of program variations to test the effect of systematically varying important features of the school-to-work model. Inevitably, the national evaluation of the STWOA will involve assessing an independent and somewhat disconnected set of State and local initiatives. Finally, only a small handful of States and localities is in a position to evaluate a mature program model that has been tested in several years of actual program experience. Most programs are. in their infancy, and many will change significantly in the next couple of years. Under these circumstances, it is imprudent to invest the political and financial resources needed to launch large-scale experimentation for the national evaluation of the STWOA. Large-scale experimentation does not seem the most practical method to achieve the broad goals of the national evaluation. Less costly quasi-experimental designs seem more sensible (see below).

Role of Small-Scale Experiments

Although it does not make sense to rely on experimentation to measure the overall effectiveness of programs funded under the STWOA, it is desirable to use small-scale experiments in some localities to test the effectiveness of selected aspects of the school-to-work model. This approach will involve persuading some local officials to allow random assignment for program evaluation. It will also require selection of a handful of questions about the school-to-work model where experimentation is both appropriate and feasible.

It is hard to persuade local authorities to administer random-assignment experiments, but it is not impossible. Program administrators have participated in dozens of random trials in the past decade in areas such as primary school class size, workforce training, welfare-to-work programs, and cashing out food stamps. Occasionally, program administrators can be persuaded to attempt experimentation when they do not have enough resources to offer a particular service to all people who are eligible to receive it. Under those circumstances, a lottery may be accepted as a fair way to decide which applicants will receive services. A lottery is simply an understandable method of random assignment, and it is also a fair way to ration scarce program services. In an experimental evaluation, students who are accepted into a school-to-work program can be compared with students who are denied enrollment because they received an unlucky draw in the lottery.

Assuming that local officials can be persuaded to administer random-assignment experiments, Federal, State, and local officials must still decide whether they can find important research questions that experiments would help to answer. It does not always make sense to conduct an experiment only because it is feasible to do so. In what circumstances is the reliability of experimental results worth the extra cost of obtaining them? The potential benefits of an experiment are clearest when the focus of study is narrow, which it is when policymakers wish to determine whether a small variation in policy would yield better results. Potential benefits are less obvious when the object of the study is to improve basic knowledge about underlying student, employer, or educator behavior-behavior that may be important across a range of schooling issues, but that has no clear implications for any single policy decision. The potential value of this kind of basic knowledge is almost impossible to determine. Not surprisingly, narrow policy experiments are much more common than social experiments aimed at improving our basic knowledge.

Randomized trials can be cost effective when most or all of the following conditions are met. We should have reasonably good evidence suggesting that a proposed program variation can be beneficial for students or employers. The proposed treatment should hold some promise of improving outcomes in comparison with the basic school-to-work approach a school will offer. In addition, there should be enough uncertainty about the benefits of the proposed alternative so that a random trial could improve the chances that the more effective (or less costly) option is eventually adopted.

An example from the congressional debate over the STWOA might be helpful. The issue of paying students for their work-based learning experiences was debated when Congress passed early versions of the STWOA. The House of Representatives passed a bill that required students to be paid, whereas the initial Senate bill contained no such provision.[4] (The final legislation seems to recommend pay for work-based learning without actually insisting on it.) Policymakers are clearly divided on the wisdom of paying students for workplace learning.

Which policy yields larger learning gains and earnings improvements to students enrolled in a school-to-work program ? Theory and available evidence do not offer conclusive evidence on this question. A randomized trial in which the two approaches are directly compared could be helpful. If there is enough uncertainty about the value of paying students for work-based learning, policymakers may find it valuable to sponsor a randomized trial testing different levels of pay (including no pay at all).

It goes without saying that an experiment should be performed only when there are no serious ethical or technical difficulties in performing it. It makes no sense to conduct an experiment if sample attrition, defective outcome data, administrative bottlenecks, or other technical problems are likely to be so severe that the findings from the experiment will be questioned.

Finally, experiments should be undertaken only when there is a reasonable chance that findings from the experiment will be used in the policy process. Because the presumed benefit of an experiment lies in the improvement of policy, there must be reasonable prospects that treatments found to be more effective are ones that State and local policymakers will eventually adopt. However, if experimental results will be ignored in policymaking, perhaps because policy decisions are made on the basis of ideology rather than cost-effectiveness, the value of an experiment will be low.

---

[4] **Office of Technology Assessment, op. cit., p. 53.**

For example, most Americans would agree that it is desirable to boost students' academic achievement as measured on standardized tests of competency and knowledge. Most also agree that it would be beneficial if students who are not bound for college entered work sooner and at higher pay after graduating from high school. A work-based learning program that achieved these goals is likely to be embraced by most voters and adopted by policymakers. It is less obvious whether voters believe it is desirable for college-bound students to learn skills in the workplace as opposed to the classroom. If considerations of ideology are likely to overshadow the issue of cost-effectiveness in reaching policy decisions, it might be best to look for an alternative topic of study. The same argument can be made regarding any evaluation research, experimental or nonexperimental. If research findings are not used to improve policy, it is hard to see how the expense of an evaluation can be justified to taxpayers.

If experimental findings are to affect policy decisions-and therefore have social value-they must produce clear, convincing evidence about the relative effectiveness of policy alternatives. In addition, they should test policies on which there is broad agreement about what a desirable outcome would look like. A variety of small-scale school-to-work experiments can satisfy these tests.

IV. An Overall Design for the National Evaluation

To answer the research questions posed by Congress in the STWOA and measure the relative effectiveness of different school-to-work models, the national evaluation should consist of the following four major parts:

A. An implementation analysis

B. An analysis of data from the STWOA management information system to measure the effectiveness of States and localities in developing and implementing school-to-work plans (see research questions 1 through 3 above)

C. A quasi-experimental analysis of selected State and local STWOA programs to determine whether services delivered under the programs have been effective in serving the needs of students and employers (see research questions 4 through 8)

D. A rigorous analysis of relative effectiveness of different school-to-work models, using either quasi-experimental comparison group methods or, preferably, one or more small-scale randomized trials

Parts A and B of this evaluation plan should be designed so that reports will be available for submission to Congress no later than September 30, 1998. Parts C and D should be designed to yield interim reports by that date, but a sensible evaluation plan would postpone the due date of the final reports until sometime in 2001 or 2002. Because the analytical problems associated with completion of parts A and B of this plan are not particularly difficult, the remainder of this report will focus on design and implementation issues connected with parts C and D.

1. Design of an Evaluation To Measure Program Effectiveness in Serving the Needs of Students and Employers

The decentralized nature of program planning and implementation under the STWOA makes it impractical to measure program effectiveness for a nationally representative sample of students and employers who participate in STWOA-funded programs. It is extremely costly to identify, much less obtain, information from a nationally representative sample of participants. It is practically impossible to identity a sample of students and employers that could serve as comparison groups for the students and employers who participate in STWOA-funded programs.

For these reasons, the only practical approach to measuring the effectiveness of programs financed under the STWOA is to draw a sample of programs in selected localities and attempt to estimate the effects of these programs using quasi-experimental comparison group methods. This approach to the evaluation raises two basic questions: How should the sample of programs be selected? How should comparison groups be drawn?

Analysts have two basic options in selecting communities and programs to be included in the study. The simplest option is to draw a random sample of school-to-work programs by identifying all public school systems in the country and then randomly selecting for inclusion in the study as many systems with operating school-to-work programs as the evaluation budget allows. Each school system could be assigned a sampling that reflects its relative size, and systems could be drawn to give large systems a much greater probability of being included in the sample. Alternatively, systems could be divided into major categories-large and small; urban, suburban, and rural; and predominantly minority or nonminority-and school systems could be randomly selected for the study so that a minimum number of systems within each category is included.

The other option is to select school systems based on the merits or attributes of the STWOA programs they run as well as the characteristics of the populations they serve and local labor market conditions. Analysts could restrict the sample to school systems that have fully implemented school-to-work systems with a track record of high student participation. They could purposefully select school systems that use a variety of school-to-work models. However, they would not attempt to enroll a nationally representative sample of school systems in the study.

Because school-to-work programs are now in their infancy in most parts of the United States, it makes most sense to use the second sampling strategy. A random selection of program sites might include a high proportion of school systems with little experience in administering school-to-work systems. Policymakers can learn less about program effectiveness from these schools than they could learn from schools that have mature programs with some track record of success. Of course, the evaluation results will not tell us how typical school-to-work programs are doing. But it would tell us something much more valuable-how well experienced program administrators can do under a mature school-to-work system. This seems to offer a better guide to what the Nation can expect to gain from full-blown implementation of alternative school-to-work programs. Parts A and B of the national evaluation will provide an excellent guide to how much progress has been made toward nationwide implementation of school-to-work systems.

Selection of Comparison Groups

Analysts also face two broad choices in selecting comparison-group members whose experiences will be contrasted with those of STWOA participants in order to estimate program effectiveness. These choices are discussed in the paragraphs below.

Selection of Students

One option is to select students from the same schools who, for some reason, failed to participate in the school's school-to-work program. Administratively, this procedure is the simplest because data gathering for both the treatment and comparison samples can be confined to the same schools. There is no reason to impose data-gathering costs on school administrators who do not operate a school-to-work program. Despite the data collection advantages, this is not a good strategy. Evaluators will face a nearly impossible task in attempting to determine what part of the outcome difference between the treatment and comparison groups is due to the effects of the STWOA program and what part is due to the measured and unmeasured background differences between the two groups that caused one group to participate and the other to decline participation in the school-to-work program.

A more promising approach is to compare the experiences of students in schools where no school-to-work program is available with students in schools where a STWOA-funded program has been established. The treatment-group sample should include students in the STWOA school, whether they actually enrolled in the school-to-work program. The goal of the evaluation should be to establish whether the opportunity to participate in a school-to-work program was effective in improving the youngsters' educational progress, job skills, and employment experiences. Subsidiary analyses would focus on the question of whether educational and employment gains in the treatment group were concentrated among those students who actually enrolled in the STWOA-funded program. However, to estimate the effectiveness of school-to-work opportunities on educational progress and employment it is necessary to compare the treatment-group members with a group that is likely to be comparable, namely, students enrolled in comparable schools where no school-to-work program is available.

Comparable schools are those in districts with similar populations and local labor market conditions. For example, if the Toledo, Ohio, school system implemented a school-to-work program and the Dayton school system did not, students from the Dayton schools might provide a good comparison group for secondary students in the Toledo schools. Of course, this analysis plan means that data must be collected from both the Toledo and Dayton schools, even though Dayton offers no school-to-work program and may not have any school administrators who are interested in cooperating with the study.

No matter how the comparison-group sample is drawn, it will be difficult to adequately represent school dropouts in the comparison sample. The STWOA envisages a system in which school dropouts are also welcome to participate. If dropouts are drawn back into school to participate in a STWOA-funded program, comparable youngsters will not be present in a sample that is drawn from the enrollment records to schools that do not have school-to-work programs. This suggests that the samples from the two schools should be drawn to represent students who have not yet reached the legal school-leaving age. The progress of these youngsters can be followed for both schools, giving investigators a good opportunity to see how dropout rates are affected and how the experiences of would-be dropouts have changed as a result of the school-to-work program.

Selection of Employers

A different strategy is needed to assess the effects of school-to-work programs on employers. Very few employers in a community actually participate in school-to-work programs, even when the local educational system has a highly successful system in which many students participate.[5] This situation makes it unfeasible to sample employers in a comparison-group community as a way to assess the experiences of employers in the treatment-group community who participate in comparison-group school-to-work programs.

In the short run, it is probably impossible to reliably answer the question "How did employers' experiences differ from what they would have been in the absence of the STWOA-funded program?" Therefore, it seems sensible to invest evaluation resources in answering other important questions. A helpful employer study would attempt to answer two sets of questions about employer involvement in school-to-work programs. First, what do participating firms think they gained from their participation? What changes in program administration would make continued participation more desirable? What changes would make the program more effective in addressing employers' needs? Second, why did nonparticipating firms fail to participate in the program? What changes in the program would increase the likelihood they would participate?

Both kinds of questions can be addressed in a survey of employers in those communities selected for inclusion in the STWOA-effectiveness study. No employer surveys would be needed in the comparison-group communities.

Randomized Trials

Part D of the overall evaluation plan involves rigorous analysis of the relative effectiveness of different school-to-work models. The proposed analysis would use either quasi-experimental comparison group methods, or preferably, one or more small-scale randomized trials. This subsection is written under the assumption that randomized trials will be used.

---

[5] For discussion of issues surrounding employer participation in school-to-work programs, see the essays in Thomas R. Bailey, ed., 1995.

The most straightforward way to conduct an experiment is to enroll students in a single school or in several schools within the same school system. In each school, students are randomly assigned to one or the other program variant and have their educational and employment progress monitored by investigators. For example, in a simple experimental test of the value of adult mentors, half the students enrolled in a school's workplace learning program would be assigned adult mentors while the remaining half would be offered access to the same school guidance system provided to students who are not enrolled in workplace learning. In a test of pay for workplace learning, students in one group would be paid for their work, whereas students in the other group would be unpaid or receive a lower rate of pay.

Conducting the randomized trial within the same school or school system has two important advantages. The data collection is very straightforward, because most information about students' educational progress should be available using an identical set of performance measures for both treatment groups. In addition, by confining the experiment to the same school or school system, the investigator does not have to be concerned about other factors that might affect student outcomes, such as differences in local labor markets or differences in other aspects of the school administration or student environment.

Of course, there is also a practical implementation problem. Students in the same school who are offered treatments that differ in their perceived value may resent being offered a less generous treatment. Administratively, it may also be difficult to keep each treatment pure, that is, uncontaminated by the alternative treatment that is available in the same school. The first problem can often be dealt with by explaining the goal of the research study. It is easier to explain an experimental test in which the favored group receives a benefit--such as adult mentoring-that a school could not afford with resources of its own. In that case, benefits are not being withheld from any student; extra benefits are being provided to a subset of students to determine whether the extra resources boost program effectiveness.

2. Implementation

Parts A and B of the overall research design should be designed for completion by September 1998. Parts C and D should provide interim reports by 1998, but the research projects should not attempt to provide final reports until 2001 or 2002. The basic reason for this delay is that it takes time for the effects of school-to-work programs to be observed and measured. Students entering the tenth grade in September 1996, who should certainly be included in the research samples, will not graduate from high school until June 1999. Therefore, it will be impossible to make reasonable assessments of the programs' effects on school graduation rates before summer 1999. These students cannot be expected to graduate from college before June 2003. Therefore, effects on college completion cannot be fully known before that year.

Although the national evaluation cannot provide definitive results on some effects of school-to-work programs before 2003, it still makes sense to begin an evaluation soon. If the evaluation is focused on effects of model programs and on differential effects of model variations, the results will be useful to State and local officials for years to come.

Data Collection

The data collection for parts A and B of the overall evaluation is reasonably straightforward. Data collection for parts C and D is much more difficult. Investigators are obliged to gather information from schools, students, and employers in order to answer 'the evaluation questions posed by Congress. To the extent possible, the evaluation design should be based on low-cost data in which the possibility of differential nonresponse in the treatment-group and comparison-group samples is minimized.

The evaluation is cheaper and easier if investigators can use information that is collected for other purposes and that covers a large, representative sample of students and employers in the treatment and comparison groups. For example, it is cheaper to rely on tests of academic performance that are administered for other purposes than it is to require students to take tests that will be used only in the evaluation. It is far cheaper to rely on earnings records available from State unemployment insurance systems than it is to interview students and school graduates to find out how much they earn.

If the evaluation relies on information that is collected for other purposes, findings from the study may be subject to less risk of bias from nonresponse error. The unemployment insurance earnings records maintained by State departments of employment security have many problems, but the problems are likely to be the same in both the treatment- and comparison-group samples. Nonresponse or missing value problems will almost certainly be identical in the two samples, which removes one source of bias when measuring treatment effects. Similarly, tests of academic achievement that are administered to all eleventh graders in a school or State will generally provide equally useful measures of student progress of both treatment- and comparison-group members.

3. Economizing on Evaluation Costs by Limiting the Scope of the Evaluation

The evaluation plan described above could require enormous resources, depending on the scale of the evaluation. The most sensible design of the final evaluation will depend on the resources Congress and the Secretaries of Education and Labor are willing to commit. At lower levels of funding, less of the plan should be undertaken. Parts A and B of the plan could be completed with a relatively modest budget, but parts C and D require larger budgets under even the most optimistic assumptions about data collection costs. Which parts of the evaluation have the greatest value to policymakers? Which have lower value?

A minimum evaluation should attempt to complete parts A and B of the overall plan. This is a minimal response to the evaluation requirements described in section 402(b) of the Act. If funding restrictions make it impossible to undertake both parts C and D of the plan, part C is the part with lower long-term value and therefore is the easiest to sacrifice.

My rationale for preferring part D over part C is straightforward. First, results from quasi-experimental statistical comparisons are always open to doubt. When the analysis is complete, policymakers will not know whether evaluators have obtained reliable estimates of program effectiveness for the programs included in the study. The comparison groups selected may be unconvincing, and the statistical analysis will inevitably be open to serious question. Second, policymakers will not have estimates for a nationally representative set of programs, only for programs that have been chosen for the study. Third, comparisons among the programs actually included in the study will not give policymakers definitive information to guide them in improving existing programs. If certain program models or program variations are shown to produce better results in the evaluation, policymakers will not know whether the superior results occurred because of superior model design or because some other aspect of the administration or environment differed between the successful and unsuccessful sites. Finally, the findings will not reflect the effects of a fully mature school-to-work program. Some of the programs included in the study will change during the course of the evaluation. Other programs have not been implemented long enough to provide definitive results.

Part D of the overall plan also has important conceptual and analytical problems. But it holds the promise of offering reliable findings about an important aspect of school-to-work programs: Can the effectiveness of these programs be improved by changing one or another feature of the basic model? For example, can learning or employment outcomes be improved by paying students for their work? Can adult mentors help students achieve higher test scores or program completion rates? No other part of the evaluation plan can provide reliable answers to these kinds of questions. If school-to-work programs are ultimately adopted throughout the country, these questions will hold the greatest interest for policymakers.

V. Summary

The evaluation plan suggested in this report consists of the following four main elements:

A. An implementation analysis

B. An analysis of data from the STWOA management information system to measure the effectiveness of States and localities in developing and implementing school-to-work plans

C. A quasi-experimental analysis of selected State and local STWOA programs to determine whether services delivered under the programs have been effective in serving the needs of students and employers

D. A rigorous analysis of relative effectiveness of different school-to-work models, using either quasi-experimental comparison group methods or, preferably, one or more small-scale randomized trials

If budget constraints make it impossible to begin all four parts of the plan, parts A and B are needed for minimal compliance with the legislative language in the STWOA. Of the other two parts of the plan, part D promises the more valuable set of results.

Net Impact of School-to-Work: Exploring Alternative

Charles Dayton
Education Consultant
Foothill Associates

I. Introduction and Background

Passage of the School-to-Work Opportunities Act (STWOA) in the spring of 1994 signaled the concern of the Federal government for the level of work preparation accorded students in this country and defined a plan for addressing this concern. The concern is well founded. Comparisons between the approaches taken in this country for preparing young people for future careers and those in many other countries leave little doubt about room for improvement here. European and Asian apprenticeship systems that direct all young people toward training that will lead to good job opportunities upon graduation from the equivalent of U.S. high schools stand in stark contrast to the lack of such systems here and the often undirected and frustrating paths that our young people wander in search of a career. This lack does untold harm to these youth, not to mention an economy that badly needs well-prepared employees.

At its broadest level, an evaluation of the STWOA needs to examine to what degree this Act has had an impact on the preparation of youth for work and the development of a more systematic approach to providing such preparation. To address these broad questions and to measure the net impact of the Act, such an evaluation needs to focus on a number of more specific questions. To provide a context for these questions and an analysis of how best to address them, this paper begins with a discussion of the central policy issues addressed by the STWOA and the questions and hypotheses to which these lead (Section II). It then moves to a discussion of the evaluation design options involved in examining these questions (Section III) and the implementation issues related to these designs (Section IV). Section V presents conclusions.

There is a long and rich history associated with the evaluation of educational programs, and I intend in this paper to draw on this history. For example, I think the separation of evaluations into phases, focusing early on implementation and only gradually shifting the focus to participant effects, is a sensible one. There are several participant-effect evaluation designs that have been employed in past evaluations of similar programs, and I will discuss these and the advantages and disadvantages of each. There are a variety of points to make with respect to sampling, instrument choice, and common implementation problems that draw on past work. None of these are revolutionary, but all are important elements of a successful evaluation.

At the same time, I intend to depart from conventional approaches to program evaluation in certain ways. First, I think the policies addressed in the STWOA are far reaching, covering a host of fundamental educational reforms with implications that go far beyond the usual school-to-work domain. Because I think these implications need to be understood for an evaluation of this Act to be well designed, I will take some time at the beginning of the paper to briefly review them. I think there are compelling reasons not to use random assignment. I suspect this will be a minority view, and I will discuss my reasons for this view. I also wish to suggest an approach not often included with such evaluations, what I have called a study of comparative efficiency, that will entail looking not just at policy and participant impacts, but costs and the relative efficiency of various approaches examined. This discussion has political ramifications, which I have tried to avoid. I think there are good reasons to seek this kind of information regardless of one's political views.

II. Conceptual and Policy Issues

The school-to-work movement is a broad one, with a number of substantive principles that bear on how young people are prepared for work and for life. An understanding of these principles and the implications for program elements that derive from them are important to an analysis of how to evaluate the effects of the STWOA. The progress being made in implementing these principles and elements will go a long way in determining the degree to which the STWOA achieves its underlying purposes. This section provides an overview of the principles and elements of the school-to-work movement. The questions and hypotheses that impact evaluation should seek to answer and test are discussed at the end of this section.

1. Collaboration

A central problem in many States and localities in the past has been the lack of collaboration among agencies and programs concerned with preparing youth for work. This begins at the Federal level with the historical split between the Department of Education's and the Department of Labor's separate and often competing programs and funding streams they have established, and their viewing in-school and out-of-school populations as somehow distinct and separate in the way they should be addressed for work preparation. This problem is being addressed currently through cooperation between these Departments, which is most welcome. But the problem is far more complicated than the historic Federal split.

In California alone, there are twenty-three different State programs directed in some way toward work preparation; and, not only do many of these programs fail to cooperate, they often do not know about each other's existence, or worse, view each other as competitors. Turf battles are rampant. Often, there is suspicion and distrust among public education institutions, private training providers, and employers. What is needed is a collaborative effort among all those with a stake in education as it affects work preparation-a coming together of viewpoints, agencies, funding streams, educational institutions, and community leaders around a common, coherent set of goals and programs. This is one of the broad intents of the STWOA, and it needs to be incorporated into examinations of its resulting impacts.

2. Pedagogy

There has been a traditional split in educational "tracks" in high schools between the "college-bound" and "non-college-bound." Those who have pursued vocational programs have been viewed in the latter category, a separate track from those pursuing academics. This practice has fostered a stigma around vocational training and, in effect, a class distinction among young people. This is unfortunate in its own right. Perhaps worse, it is unfortunate from a pedagogical viewpoint. As cognitive scientists have demonstrated, learning is most effective when done in a meaningful context, and careers provide perhaps the most comprehensive and meaningful context available to illustrate to students the relevance of their academic subjects: English and language arts, math, science, and history and social studies. By joining academic and vocational topics through curricular integration, students can be shown how what they are learning is meaningful to what they will encounter as adults, thus improving their motivation and removing the stigma associated with vocational training.

A second feature of needed pedagogical reform is the articulation of education across various grade levels. As school districts have grown in size and split into multiple campuses, often there is little contact among elementary, middle, and high school levels. Although some dedicated teachers take it upon themselves to understand the preparation needed for the next level, this is too rarely systematic. Even more troubling, links within the high school curriculum, other than in terms of preparation for four-year college programs, has been even rarer. Given the fact that less than one-fourth of each cohort of students completes a baccalaureate degree,[1] a majority of students are unfocused in their studies and ill prepared to transition to a postsecondary option. Indeed, 88 percent of those not planning on attending a four-year college have no curricular focus in high school.[2] The tech-prep movement has addressed this problem in recent years, establishing links between secondary and postsecondary courses and programs. The movement toward structuring high school curriculum around career majors or career paths has fit well with Tech Prep and has also helped. These are additional important elements of pedagogical reform.

3. Serving All Students

A related theme of the Act is the need to develop policies and programs directed to *all* students. This pertains to the issue of class distinctions and vocational stigmas discussed above. Until career preparation is regarded as important for all students, such stigmas will remain. But it also relates to the issue of pedagogy. Students planning on attending college can also benefit from seeing the relationship between academic subjects and various careers. They will work one day also; the only difference is that their work lives will begin after college rather than before. While less than one-fourth of each cohort graduates from a four-year college, it is nevertheless important that this segment of students be addressed in a school-to-work program as well as those who will not hold a bachelor's degree.

---

[1] Mini Digest of Education Statistics, 1994, p. 40.

[2] National Center for Research in Vocational Education, 1994, p. 2.

This issue also extends to many other populations. Gender is important. Although women now have on average slightly more years of education than men, women continue to earn about seventy cents for each dollar men earn. Race and ethnicity also affect the quality of education available to many students and the careers open to them. In California, the majority of kindergarten through twelfth grade students now come from minority groups, yet such inequities remain. Language is often a part of this issue; in California over twenty percent of kindergarten through twelfth grade students speak English as a second language.[3] This can be an advantage in a global marketplace, but must be viewed as such in training programs for that to occur. There are also many other subcategories of students to be considered. One of the objectives of the STWOA is to bring fairness and equality to work preparation, and this issue needs to be incorporated in the evaluation.

4. Technology and the High-Performance Workplace

Technology is a central driving force in most industries, especially those with strong growth and increasing employment potential for young people. Yet schools are usually far behind the curve of technological change. It makes little sense to prepare students for tomorrow's work setting on yesterday's equipment, but that is what often happens. Part of the change that work preparation needs involves ways of bringing training technology to the level used in industry. This can be done by bringing improved technology (and heightened awareness of its importance) to schools and by bringing more students to high-tech workplaces. These are important outcomes implied in the STWOA.

Related to this is the fact that too often, students are made aware of career options and provided with work preparation, for workplaces as they have been structured in the past rather than those that will predominate in the future. As the Secretary's Commission on Achieving Necessary Skills (SCANS) reports illustrate, the skills required for high-performance workplaces are in many ways different from those of the past.[4] Production in high-performance workplaces is flexible, customized, and decentralized rather than centrally controlled around mass production techniques. Workers all become part of quality control, working in teams requiring cooperation, continual upgrading of skills, and the ability to perform a variety of tasks rather than as lone operators performing the same limited task repeatedly over long periods of time. Thus, the skills that workers need are becoming more complex, at a higher level and more challenging. Further, with widespread downsizing and the trend toward leaner corporate structures, such skills are ever more central to job security. Both the improved awareness and use of technology in schools and these changes in the workplace need to be reflected in policies and programs developed through the STWOA, and the evaluation should examine the degree to which they are.

---

[3] Policy Analysis for California Education, 1995, p. 97.

[4] Department of Labor, 1991, p. 3.

5. Teaching All Aspects of the Industry

In the past, vocational education at the secondary level has often tried to prepare students for specific jobs. When workplaces were relatively stable and skill levels relatively low, this made sense. Today, it makes sense less and less. Given the changing nature of the workplace, at the secondary level students increasingly need to understand the broad features of an industry rather than be trained for specific jobs within it. They need to learn skills that will be transferable from one company and job to another so they can adapt to change. Since most new jobs will require some training beyond high school, a more specific focus can occur during postsecondary training. The STWOA discusses this need and some of the broad industry features to be addressed: planning; management; finances; technical and production skills; underlying principles of technology; labor and community issues; and health, safety, and environmental issues. In examining the policies and programs promoted through the STWOA, one needs to examine whether there is a focus away from traditional job-specific vocational training and toward this broader focus on all aspects of the industry.

6. Standards, Assessment, and Credentialing

One of the themes of the STWOA is the need to move toward a system of work preparation based on clearly defined standards that describe what young people should know and be able to do. The SCANS reports go a long way toward defining such skills at a generic level. There are also many skills relevant to particular industries. The twenty-two industry-specific, standards-setting projects being completed, funded jointly by Education and Labor, offer a good start in this direction.

GOALS 2000 offers a similar perspective and function relative to academic standards, as does the New Standards Project led by the National Center on Education and the Economy.[5]

Such a reliance on clearly defined standards leads directly to the issue of assessment. Much of assessment in traditional academic disciplines tends to center around norms-referenced testing, which, when graded on a curve, leads to comparisons of students against each other rather than how well all are meeting defined standards. Further, such testing measures primarily how well students can memorize and guess rather than their ability to seek out information, analyze situations, and apply knowledge to real-world problems. Assessment that measures how well students can perform in these ways, often referred to as "authentic assessment," holds far more meaning in terms of workplace preparation.

Finally, just how and how well students are recognized for their abilities are important issues. High school transcripts hold meaning for college entrance requirements, for which they were originally designed. They hold little meaning for employment; few employers even request them in making hiring decisions. Thus, for the majority of students who do not attend a four-year college, some form of recognition of their abilities and workplace readiness needs to be evolved that has meaning among employers. This is important not only to give students recognition but to provide an incentive for them to take their high school education seriously.

A variety of approaches has been suggested in this regard, perhaps leading among them are the Certificates of Initial Mastery and Technical and Professional Certificates in America's *Choice*.[6] Alternatively, some States are examining ways to make the high school diploma more meaningful to employers. Regardless, this is a problem that needs to be addressed in attempts to improve school-to-work programs. Thus, the degree to which progress is being made regarding standards, assessment, and credentialing is another important issue to examine in this evaluation.

---

[5] Learning Research and Development Center, 1995.

[6] National Center on Education and the Economy, 1990, pp. 71, 77.

7. Fostering Systemic Change in Schools

Many have charged that school reform efforts have been plagued by "programitis." Instituting a variety of often unrelated or overlapping programs for one subset or another of students, while perhaps well intentioned and better than nothing, is not the same as undertaking fundamental reform of the central structures of education in a way that will affect all students. While a number of the elements listed below have been touched on previously, the point in restating them here is to emphasize how they need to become part of systemic change, not isolated add-ons here and there. There are important ways in which such systemic change is needed:

- Movement away from separate, discipline-based instruction toward contextual learning that demonstrates relatedness among different subjects, especially between academic subjects and various industries and careers. Such a change implies the use of team teaching, reformed schedules, such as use of block scheduling, and often a merging of administrative and teaching roles.

- Development and use of well-defined standards, authentic assessment, and meaningful credentials.

- A collaborative approach in which schools become community-based institutions in the sense that all those with a stake in education-employers, parents, labor unions, community members, and students themselves-have a role in defining their goals and contributing to their success.

- Instruction and assessment that emphasize the importance of cooperative learning, so that students learn to work effectively in teams and view classmates as "coworkers" rather than "competitors." Heterogeneous grouping and peer tutoring can be part of this element.

- Breaking of large, impersonal high schools into smaller units, "schools-within-schools," designed to provide more personal contact with students, each with its own team of teachers and often focused around a common theme.

- Use of career majors, often called "career paths" or "career path clusters," to frame the last years of high school, providing every student with a means of focusing the curriculum around a real-world, postsecondary goal and learning the relationship between academic subjects and a variety of careers.

- Articulation of instruction and coursework across different levels of education. This can occur within the kindergarten through twelfth grade system, so that elementary instruction relates more clearly to that at the middle school, and middle school to high school. Perhaps most important is relating high school to postsecondary levels, as defined in the tech-prep movement, so that students graduate with clearly defined postsecondary goals, knowledge of educational opportunities available to pursue those goals, and a start in the next level of education to provide a foothold.

This list, while not exhaustive, represents a substantial set of educational reform elements. All are important to the development and success of a school-to-work system and need to be examined as part of assessing progress at the State and local levels.

8. Employer Involvement

Unlike the case of our European and Asian competitors, there is little history of collaboration between public education institutions and employers in the United States concerning the preparation of young people for work. In apprenticeship systems common to many other countries, there is a close working relationship between educators and the employers and workers in the companies to which students will be moving upon graduation from "high school." Not only do these partners agree on the standards and methods of training, they cooperate in the delivery of this training. Thus, students move in an almost seamless fashion from school to a productive role in a company.

In contrast, employers in the United States have generally not regarded it as part of their responsibility to help prepare young people for work, except in terms of specific jobs in their companies or agencies after employees are hired. With the emphasis in the STWOA on work-based learning, this raises an important question concerning the degree to which employers will be willing to contribute the time and energy needed to develop and support such a system that will benefit the general good but perhaps not serve their particular needs in any immediate fashion. With the emphasis on short-term profits, competition, and the trend toward downsizing in this country in the last decade, this issue is particularly relevant. Unless employers play the needed role, the intended partnerships between education and business will be only shells, unable to fulfill the roles necessary to accomplish the job. This question needs to be closely examined.

9. Establishing Work- and Community-Based Learning

Kindergarten through twelfth grade education is something that in most instances takes place almost exclusively on the grounds of school campuses. In order for students, especially high school students, to learn about the world of work and the skills required in the workplace, experiences in work settings can provide a rich complement to school-based learning. Thus, an important element of the STWOA is the joining of workplace learning experiences with those based on high school campuses. Such experiences can occur at several levels of intensity, from simple field trips or one-on-one job shadowing, to longer term mentoring by an employee of a student and periods of either unpaid or paid supervised employment. Workplace learning has been found to be most effective if it is carefully planned, with clearly defined learning goals, and tied to classroom instruction. Thus, a series of connecting activities between work and school activities is also important, including training of both teachers and job supervisors in the purposes and methods of workplace learning.

One of the challenges faced by attempts to place large numbers of students in work settings is the shortage of settings that employers are willing to make available for such purposes. As discussed, employers lack a history in this country of participation in work-preparation education except as it directly benefits their own employees and company. An alternative venue for such real-world learning is community service projects. While these are not usually paid, they can provide students with experiences that show them the requirements of work, how knowledge can be applied to real-life settings and problems, and the importance of civic responsibility. The degree to which the STWOA programs develop work- and community-based learning components is another important focus of this evaluation.

10. Professional Development, Resources, and Technical Assistance

The elements that have been discussed represent a substantial set of changes needed in the way we prepare young people for work and in the way both teachers and employers operate. It will not happen without substantial resources devoted to this effort. Professional development is needed in all of these areas so that those expected to deliver these components understand their purposes and how to effectively deliver them. There are many sources to draw on for such training and technical assistance. A central issue in how well STWOA efforts proceed in each State and locality is the degree to which these resources are marshaled effectively and brought to bear to realize the achievement of each of these elements. This, too, is an important issue for this evaluation to examine.

Summary of Reform Principles and Program Elements

These ten principles of school-to-work educational reform and their related implications for program structures represent substantial redirection of the Nation's education and training. Such changes require efforts at several levels. The Federal Government has a role: to provide direction and leadership, establish standards, and supply whatever resources are feasible to support changes in these directions at State and local levels. The STWOA represents a substantial step forward with the first of these roles. Effective implementation of this Act is the next step.

States, too, have important roles. They need to establish their own plans and directions under the Act and offer leadership and support to regional and local efforts within their States. The degree to which they effectively translate the Act into State-level policies and regional and local programs is crucial in determining the impact the Act will have. The State can also set examples of how to proceed in some of the policy directions, such as bringing various agencies and funding streams together into a collaborative effort.

Finally, regional and local efforts are critical. Here, the issues are less related to policy development than to translation of policies into effective actions that reform education in the ways policies point. This means redirecting educational institutions and establishing school-to-work programs that positively affect young people. Ultimately, the measure of success of the STWOA will be the success of the Nation's youth in their school-to-work preparation and subsequent careers.

Questions To Be Addressed and Hypotheses To Be Tested

Given this set of intended policy and program impacts, what are the central questions to be addressed in this evaluation? First, they exist at three levels: Federal, State, and regional and local. We need to examine what is going on at each level and compare this activity with what was intended in the Act. Second, we need to look at three aspects of such activity: (1) the quality of implementation, (2) the policy impacts, and (3) the effects on participants. The first can be done at all three levels, the second primarily at the State and regional and local levels (the Act defines Federal policy), and the third primarily at the regional and local levels. The broad, underlying questions that such an investigation should address are essentially the following:

1. To what degree are the policy elements of the STWOA being achieved? That is, to what degree are the ten elements discussed earlier in evidence and more in evidence as a result of efforts conducted via the STWOA?

2. To what degree are STWOA programs having the desired effects on the work preparation of those young people involved in them?

In examining these questions, there will undoubtedly be variations in the emphasis given to alternative policy elements in various States and localities. There will also be variations in the way the policies are translated into educational reforms and programs as they directly affect students. These deviations can represent natural evaluation design variations in policy directions and program elements. The resulting effects associated with these variations can then provide information about what works best and where. Ultimately, the hypothesis to which all these efforts relate is the following:

- In those sites where the policy and program elements of the STWOA are being more fully and successfully implemented, greater success will be found in the preparation of young people for work.

Another set of considerations to be made in this analysis leads to a different kind of hypothesis. The above hypothesis is based on a certain underlying assumption: that there are some approaches to implementing this federally determined set of policies and programs that can be considered worth the cost. But not all would likely accept this assumption. Thus, another, perhaps more fundamental question in this evaluation might be whether there are any ***sufficiently*** effective approaches to the implementation of the STWOA's policy and program intents to justify its expenditures. This question leads to a different line of inquiry. Some of the questions it points to are the following:

- What would represent sufficient effectiveness in this arena?
- What is an acceptable amount to spend on such efforts?
- When is the cost-benefit ratio sufficient to justify Federal action in this realm?
- How do the efforts of the STWOA compare with other efforts conducted outside this Act?
- How can future efforts be structured to ensure that only the most efficient approaches are employed?

These questions lead to a hypothesis that might be stated as follows:

- It is possible under certain conditions and with certain approaches to implement the policies and programs of the STWOA in a way that is sufficiently cost efficient to justify the Act's expenditures.

The evaluation then becomes directed not only toward establishing what has been achieved in terms of policy and program impacts, but toward examining these against what are deemed acceptable levels of achievement, as well as acceptable levels of Federal support. Granted, these questions are to some extent political questions for which no evaluation can provide final answers. But an evaluation can be directed toward gathering information useful in addressing such questions. One implication, for example, would entail gathering data on other comparable efforts and examining and comparing cost-benefit ratios as well as simple outcome data.

III. Design Issues

As with any evaluation, this one needs to be broken down into its various elements. There are several elements. While there are various ways to categorize these as they relate to the first hypothesis discussed above, I have divided them into three basic aspects: implementation, policy impacts, and participant effects. A fourth aspect that relates to the second hypothesis I have called "comparative efficiency."

1. Implementation

Some of the questions that need to be looked at pertain not to results but inputs: what has been done to effect the hoped-for changes. This is often called "formative evaluation." These are questions that relate to how well the Act is being implemented. We cannot simply assume that everything that was intended to be implemented under the Act is, in fact, happening, even at the most basic level, such as whether information about the Act has reached the decisionmakers at the State and local levels or whether the funds intended to be distributed under the Act have been distributed. Examples of questions in this category include the following:

- What mechanisms have been effected to implement the Act? What have the agencies assigned responsibility for this at the Federal level, in fact, done? How much funding has been distributed and where? How much information has been distributed and where? How are the mechanisms designed for these purposes working? What have the problems been and how might these be addressed?

- How are State governments handling their roles under the Act? What mechanisms have they put in place to handle the policy and funding directives of the Act? What problems are they having? What features of the Act are they finding easiest and hardest to implement?

- What roles have regional and local collaborative bodies taken on in establishing policies and distributing funding at this level? What is working well or poorly at this level?

- What sort of programs that directly affect students are being established under the Act? How well are they being implemented? Funded? What have been the successes and problems in this regard?

The answers to these questions generally can help program managers at the various levels refine their procedures so that implementation works more smoothly and effectively in the future.

2. Policy Impacts

This category of evaluation addresses the question of what changes have taken place in the policies and programs targeted by the Act as a result of the implementation that has occurred. In a sense, this is an intermediary level of evaluation, because it addresses outcomes that are assumed will benefit participants. They are not measures of actual participant change. Nevertheless, they represent important outcomes in their own right. The questions in this realm are essentially ones that address whether the intended policies have been effected. For example:

- How have policies at the Federal level changed as a result of the Act? What are Federal agencies doing differently as a result of the Act?

- What are State governments doing differently as they affect policies and programs in this arena? For example, is there greater collaboration and more emphasis on systemic change? How has the decisionmaking process been affected? How are funds flowing differently?

- Are State departments of education reflecting the policies of the STWOA, as discussed earlier, in their operations and directives to school districts?

- How has decisionmaking at the regional and local levels changed? Have previously existing program administrations and funding streams, in fact, come together as hoped? Is collaboration growing? Are efforts more systematic than previously?

- Are programs that affect students more responsive to the principles that underlie the Act? For example, is pedagogy moving toward academic-vocational integration and cross-level articulation? Is there a greater awareness of the importance of technology in these programs? Is employer involvement increasing? Are programs focusing on "all aspects of the industry" rather than specific vocational training? Are more students experiencing workplace learning? Are industry standards more in evidence and assessments more authentic? Are schools undergoing systemic change in the ways discussed above?

Answers to these questions determine whether the policies that were intended under the Act are in fact being carried out at each level-Federal, State, and regional and local. They answer the first broad, underlying question posed above, "To what degree are the policy elements of the STWOA being achieved?"

3. Participant Effects

This leads to the third category of evaluation that relates to participant changes. In the realm of education and training, the ultimate bottom line is how those undergoing the treatment are being affected. Even if the Act is being well implemented and the policies and directions defined in the Act are being effected at the Federal, State, and local levels, the ultimate question is whether program participants are benefiting. The first step in designing an evaluation in this category is identifying the measures to examine. This entails finding indicators of desired program participant behavior that can be measured, so that improvement over time can be determined. Examples of such indicators that might be considered for in-school participants include the following:

- Attendance rates
- Credits earned toward graduation
- Retention in school or the program through some predefined point (e.g., a school year, graduation, and end of program treatment)
- Grade point averages (GPAs)
- Test scores, and assessment results (e.g., proficiency tests and portfolio results)
- Graduation rates

While in-school results can be effective measures of program success, in the realm of school-to-work probably a more important category of participant outcomes is that related to postsecondary results. After all, these are school-to-work programs, so that the school phase of the experience represents the first part, not the last. Examples of participant measures in this realm might be the following:

- Enrollment in a postsecondary training program

- Completion of a postsecondary certificate or degree
- Part- or full-time employment
- Length of employment
- Measures of employment quality (e.g., job title, wage rate, and opportunity for advancement)

Answers to these kinds of questions address the second broad, underlying question posed above, "To what degree are STWOA programs having the desired effects on the work preparation of those young people involved in them?" Together with the above two categories of information, they allow one to test the first hypothesis, "In those sites where the policy and program elements of the STWOA are being more fully and successfully implemented, greater success will be found in the preparation of young people for work."

4. Comparative Efficiency

Efficiency refers to both effectiveness and cost. That which is most efficient provides the greatest effect at the least cost. It relates to broader political considerations than simply an examination of how the Act is doing in its own right. An evaluation focused in this direction would compare results across various settings and program approaches. Some of this information might be derived from comparisons among States or among programs within States. It might also include examinations of other countries' efforts, comparing what is being attempted here and elsewhere, as well as what is being achieved.

This approach would also require gathering cost data. It would be necessary to know the annual cost per student for each program evaluated and from what sources such funds derive. This would allow an inspection of what State and local resources have been leveraged through the Act and to what degree Federal support versus other sources plays a role in school-to-work efforts.

Another element that would be included in this type of an evaluation would be an analysis of the optimum conditions for success at both the policy and program levels. One of the results might be to define a set of conditions under which success at the policy or program level is most likely to occur and, conversely, when it is least likely to occur. This could provide guidance in the future for more selective and efficient use of limited Federal resources.

By comparing both effects and costs across various STWOA efforts, and between these and non-STWOA efforts, one could produce comparative cost-benefit analyses, leading to conclusions about what approaches are most efficient. Information of this sort might help to answer the question of whether Federal action is justified by determining what level of Federal support can reasonably be expected to produce what outcomes and under what conditions. It might result in a set of guidelines defining the optimum conditions and direction for Federal action. Ultimately, the purpose would be to illustrate how Federal resources can be used more efficiently to ensure that limited resources produce the greatest possible effects. Information in this category would allow an answer to the second hypothesis posed above, "It is possible under certain conditions and with certain approaches to implement the policies and programs of the STWOA in a way that is sufficiently cost efficient to justify the Act's expenditures."

5. Instruments

The most sensible form of data collection tends to be defined by the category of evaluation involved. Each of the above categories entails certain kinds of instruments and data. For example, in the first two categories, implementation and policy impact, information usually comes from a combination of the following:

- Document reviews, such as for each of the State's plans under the STWOA, regional and local collaborative plans, and program designs and reports.
- Interviews with key actors at the Federal level, State legislators, State agency staff, regional and local collaborative representatives, and program managers.
- Structured questionnaires, particularly in categories where there are common positions from one State to another and it is important to have comparable data. These will focus on different topics in different situations. For example, the State-level ones will focus mostly on management and policy changes, while the program-level ones will get into details of the reforms as they directly affect participants.

In the realm of participant changes, a somewhat different set of instruments will be most sensible. Some information in this category can come from interviews and questionnaires (with program administrators, teachers, employers, and participants). However, added to this will probably be some use of observations, so that evaluators can form their own opinions about the breadth and depth of intended reforms at the participant level. There also will be considerable data to be collected from program and student records (such as attendance rates, credits earned, GPAs, graduation rates, and postsecondary education and employment plans).

At this level, it will also be useful to conduct followup surveys of high school and program graduates to learn how they are faring after leaving the training provided. These are the ultimate measures of impact at this level. Such surveys need to include information on both postsecondary education and employment. The type and level of education reached needs to be known, as well as the type and level of employment.

Analyses of comparative efficiency will require collection of cost data from various records: State expenditures per locale, local expenditures per program, and program expenditures per participant. Expenditures can be itemized in terms of various sources: Federal, State, and local. In the last category, much of the support will probably be in-kind support; and given the experience with California's Partnership Academies, the largest category that is both needed and contributed is people's time. Formulas can be developed for estimating the worth of time contributed by school personnel: teachers, counselors, administrators, and aides. The same is true for employer support, such as business speakers, field trip hosts, mentors, and workplace learning supervisors. Other categories can also be included, such as transportation, equipment, facilities, instructional materials, and supplies.

There are, of course, many issues of instrument design to be considered. It seems premature to consider this level of concern here, but it is important that sensible rules of instrument design and use be followed. Response burden is one important consideration. Instruments should be kept as short as possible and gather only information of genuine use. Obtrusiveness is a second important issue. Where equivalent data can be obtained in more than one way, the way that least affects program operators and participants is generally preferable. A third issue is planning in advance for the intended statistical tests so that data are collected in the form that best lends itself to the intended tests.

6. Sampling

Sampling is always an issue in an evaluation of a substantial Federal program. It is neither feasible nor sensible to examine every instance of every policy and program established under the Act. A well-focused study can reveal fully as much information as an unfocused one, and it costs much less. Different aspects of the evaluation argue for different approaches to sampling. Implementation, for example, can and probably should be examined everywhere. Much of this information should be available with a relatively limited effort, such as through existing reports from localities to States and States to the Federal Government.

Examining policy impacts will require somewhat more indepth mining of data, and here I would suggest a two-pronged approach. Certain forms of basic information should be sought from all States. It should be known, for example, which States focused to what degree on each of the ten elements discussed earlier, how they effected such policies at the State level, and what the successes and problems have been. However, more indepth questions in this arena might be approached by identifying a few States that represent good examples of one approach or another for more detailed data gathering. Identifying such States would require establishing a set of criteria that would ensure capturing the broad cross section of approaches that exist. Such criteria might include geography; State size; preexisting conditions as they relate to school-to-work; level of funding, effort, and implementation; policy emphasis; regional and local implementation; and program designs as they affect students.

It is in the third realm, participant effects, that sampling is the most serious issue. Student outcome evaluations are expensive, especially if they involve tracking participants over several years, which this study needs to do. Thus, I would suggest a serious sampling strategy at this level. The States chosen for indepth policy impact analysis should be those from which this sample is selected. Within these selected States, a small group of local sites and programs should be picked for indepth, long-term study. Again, criteria should be established for ensuring a good cross-section of programs. Without knowing the resources available for the evaluation, it is difficult to attach meaningful numbers to this discussion, but the fewer sites that will provide adequate data the better. As a broad estimate, I would probably aim for ten States to include at the indepth level and perhaps two to three local sites within each. Within each local site, enough participants need to be included to allow sensible use of statistical tests. Such testing needs to be by cohort, so sample sizes need to be determined by cohort. Again, this is a somewhat arbitrary number. I would aim for perhaps fifty participants per cohort (one hundred would be preferable but may be harder to find and will increase the cost) and would suggest following two or three cohorts per site (i.e., identifying new cohorts for two or three years).

7. Comparisons Across States and Localities

States will have taken many different approaches to the implementation of the STWOA. The first issue here is to learn just what these variations are. Much of this information can come from the implementation part of the evaluation. Comparison can then be made across States on the following several dimensions:

- The degree to which the Act has been implemented
- How state-level operations have been affected
- How regional and local operations have been affected
- How programs have been designed and structured
- How participants have been affected
- Costs at each of the levels

Such comparisons can be across each of these dimensions, both among STWOA efforts and between these efforts and non-STWOA ones.

8. Design Options-Participant Effects

A great deal has been written and debated about the best evaluation designs for gauging participant changes in educational programs. There are a number of alternatives in this respect. Perhaps the three of most interest are (1) time series, (2) comparison group, and (3) control group. The first of these entails measuring the change in performance at various points over some period of time for a given group of participants. Since no comparison students are involved, this is a nonexperimental design. The second alternative adds to the first a group of nonparticipant students who are selected to be as similar as possible to the participants and follows both groups over some period of time. This is a quasi-experimental design. The third and true experimental design entails choosing the "treatment" (participant) and "control" groups from one general pool of students through a system of random selection. Each has advantages and disadvantages, and none of the three is perfect.

Time Series

The time series or predesign and postdesign is the least expensive and easiest to implement of the three options. First, it entails only those students involved in the program, so the number of students to be followed is smallest. Second, it is the least obtrusive, since the program generally must operate the way it would without an evaluation as far as its design and structure go; the only addition is collection of certain kinds of data, much of which are available from existing records. Third, the data analysis is easiest, since no statistical tests are required to determine differences in performance between alternative groups, although pretests and posttests are needed.

The biggest difficulty with this design is that it also yields the least conclusive information. Without any comparison data, it can always be argued that any changes in performance, even if substantial, are due to simple maturation. Seniors generally perform better than freshman and more mature students better than less mature ones. So, how can one know if the improvements found in performance over time are attributable to the program? There are ways to ameliorate this difficulty. Sometimes, broad existing measures can yield meaningful comparisons. For example, the California Partnership Academies, which serve almost 5,000 mostly at-risk students in forty-five high schools in California, each year show a three-year high school dropout rate (grades ten to twelve) of approximately eight percent. The statewide dropout rate for those same three years is nearly sixteen percent. Thus, with no direct comparison data, it is possible to show considerable impact from the Academies' programs. However, finding meaningful, broad comparison data of this sort is often difficult.

Comparison Group

This approach, too, has several advantages. While more elaborate and expensive than the time series design, it is simpler and less expensive than the experimental design. It is also less obtrusive than the experimental design, because, again, program managers can generally operate as usual; they do not have to alter how they select students to be involved. The addition of comparison groups to the time series design also considerably strengthens one's ability to attribute any improvements in student performance to the program rather than simple maturation, since these improvements will be measured against another group that is going through the same time period but without the treatment.

However, this design, too, has its weaknesses. Identifying students to be in the comparison groups is a complex and difficult process. To be useful for comparison purposes, they must be carefully matched. For example, in the California Partnership Academies evaluation conducted between 1985 and 1988, comparison students were matched on high school, grade level, age, gender, race, ethnicity, and on several measures of preprogram performance (in this case, ninth-grade performance), including attendance, credits, grades, and standardized test scores. Gathering all this information on students can take a lot of work.

Another problem is sensitivity on the part of schools and districts about providing data on comparison groups purely for purposes of research. Sometimes such data are only available on an anonymous basis (i.e., if the students involved do not know they are part of a comparison group). This permits availability of certain kinds of data (e.g., those available in existing records) but not others (e.g., those from a questionnaire).

Another problem with this design, perhaps the most damaging, is the fact that even if well done, it can still leave questions about attribution of differences between participants and comparisons. This is because comparison groups are generally selected from populations of students that did not apply for a program, and the one factor that cannot be matched is motivation. It can always be argued by a skeptic that even large differences in performance that are statistically significant between program and comparison group students are due to the fact that the program students were more motivated from the start.

Control Group

Because of these kinds of difficulties, the predominant design usually proposed in this kind of evaluation is the true experimental one involving randomly assigned treatment and control students. These must all come from pools of students who want to be in the treatment, eliminating motivational differences. Which students end up in which category is then determined by a random selection process. If the groups are of sufficient size, this process eliminates any question about whether differences in performance between them are due to the program influence; theoretically, it is the only difference between them. The random process is intended to ensure equivalent groups in every respect. It is unarguably the strongest design from a statistical viewpoint.

However, this design, too, has its drawbacks. It is the most intrusive of the three options. It requires programs to recruit more applicants than can be served, ensuring the disappointment of those assigned to the control group. It has been known to trigger attempts to sabotage programs because of resentment from this source. Teachers and administrators often resent the loss of control they suffer in the selection of participants. Many school districts have established policies preventing use of this design because of the ethical questions it raises. Perhaps the clearest way to state these concerns is that when students become part of an experimentally designed evaluation, their futures are being determined not by their needs but by those of research. A counterargument is that until we know more about what really works, we are failing to serve all students as well as we could; and if a program is overenrolled, random assignment is as fair a system as any other for deciding who gets in and who does not. In any event, it is widely argued for and more often used in educational evaluations.

While there are unquestionable advantages to the true experimental design from a statistical standpoint, and this is an important consideration, I am not an advocate of this approach. I find the ethical concerns it raises to be real ones. Many districts that have examined this issue have ruled that this design cannot be employed in the district because of these issues. The point at which one decides to use a control group design is the point at which one has concluded that research needs should drive educational choices, instead of the other way around. Any time a decision about the educational future of a student is made using a table of random numbers, a research-caused injustice is possible. Human judgment in such matters is not infallible either. But at least it is not founded on the principle that half the students in a pool under consideration for a treatment will be arbitrarily put into a nontreatment control group so that we can use them for purposes of study, with the presumptive hope that they will perform less well in the future on the dimensions under study (if we hope the treatment works).

Students often resent the fact that after applying and qualifying for a program they want to enroll in, they must instead serve as a control in a research study, while various friends and acquaintances are allowed to enjoy whatever benefits the program may provide. Although they may or may not understand the reasoning behind the use of control groups, few come away happy and some clearly feel mistreated. Given the fact that school-to-work programs often serve students who have not felt particularly well treated by the education system in the first place, I think it is fair to ask whether, as researchers, we ought to be imposing one more seeming injustice on such youth.

The problems of intrusion are also real ones with an experimental design. Many teachers and administrators resent the loss of control it provides for them to make judgments about which set of students is most likely to benefit from a program. Some teachers who have worked in programs evaluated in this way have felt that the pool of students that ended up in the program was less well fitted to the treatment because of the extra recruitment efforts necessary to increase the size of the pool to accommodate both treatment and controls. I think one also needs to question the veracity of the data obtained from controls on questionnaires, given what may be their resentment toward the program and therefore their potential response bias.

Finally, there is the problem that in educational research there is no such thing as a "control" group, in the same sense the term is used in other types of research. This is not medical research, where a placebo can be given, a true nontreatment, and no other conditions are varied. Every control student will, in fact, be undergoing some alternative treatment, not a nontreatment. The alternatives themselves will vary and vary in effectiveness. Thus, any design that employs comparisons in fact is a far more complicated design than it appears on the surface.

Related to this is the problem that while the theory behind an experimental design is that any variations will sort themselves equally into treatment and nontreatment groups, there are, in fact, certain potential biasing factors that may work *across* either of the groups, other than the treatment under study. Many educators feel that most of the variation in any program is due to teacher quality, and that one of the reasons it is so hard to replicate successful programs is because teachers cannot be cloned. The same might be said of administrators or counselors. Other factors that might bias one group over another include equipment, facilities, or educational materials that may not be thought of as part of the treatment or nontreatment.

The point is that research in education is not, and by its nature cannot be, an exact science. We need to use a variety of measures in assessing any impact and to consider teacher and administrator judgments along with more objective "outcome" data. When a variety of measures comes together to suggest a common finding, only then can we begin to trust it. Trusting an individual statistical indicator derived from any design is risky, be it experimental or otherwise.

I realize this leaves open the question of the best participant outcome design to use in this study. Because of the issues discussed above, I have concerns about all the options. It is important that all their limitations be recognized. However, given all the factors to consider, I would choose the comparison group design. It surely is statistically imperfect. But if cautiously employed, which implies careful matching on a number of dimensions, including preprogram performance, with large enough cohorts, in my judgment, it is relatively reliable. It is less obtrusive than the experimental design and avoids its ethical problems. It should be used in combination with other approaches that incorporate human observation and judgment. It is far from perfect but so is educational research itself.

I would make one exception to my uneasiness with the experimental design. If there are two or more alternative treatments under study, each of which is thought likely to produce positive results, I find no objection to using random assignment to place students in such alternatives. In this instance, research needs are not overriding educational ones as they impact individual students.

The question was also presented at what point in time might random assignments be applied. I think the best answer is at the beginning of the treatment, however this is defined. Because in a school-to-work program it seems fair to assume this will be at the beginning of whatever the school portion of the treatment is, this time would be the best time to establish control (or comparison) groups. Any subsequent use of the technique would be restricted by the fact that students beyond this point are rarely part of a consistently defined group. Workplace learning assignments will presumably entail placing students in a variety of settings. Perhaps, if at some point beyond the initial one there were such group differentiation in treatment, it might be possible to define control or comparison subgroups (e.g., if one group that experienced the school program goes on to a workplace learning experience and another does not). But given the fact that such programs are attempting to make systemic changes, I suspect (and hope) such instances will be rare.

Iv. Implementation Issues

1. Common Problems

I would like to begin this section with a discussion of some of the common evaluation implementation problems involved in the use of any design. Regardless of the design, there are dangers that can weaken or destroy any findings if they are not carefully considered and avoided. One of these is attrition. It does not matter what the design is; if there is differential attrition between the treatment and comparison or control groups (meaning there are more dropouts in one than the other), this may bias any comparisons between them. This concern can be eliminated if every original member of both groups is tracked and followed throughout the period of the evaluation, regardless of where they go or what they do. However, this is expensive and difficult.

A second problem of any design is that the most common finding in such evaluations is that of no statistically significant differences between groups. It is relatively rare to find clear-cut performance differences between treatment and comparison or control groups. This is not so surprising if one considers the multiplicity of influences in any person's life and the relatively small portion of such influences an educational program is likely to have. Thus, evaluation sponsors are often in the unhappy position of having spent large sums of money to learn very little.

A third weakness of any design is determining what it is about the treatment that causes differences in performance. Most treatments involve a variety of elements. Simply finding clear differences between treatments and comparisons or controls, rare as they may be, still reveals little about what elements of the treatment made the difference. Since the 198588 California Partnership Academies evaluation found a fairly clear-cut pattern of statistically significant differences between program and comparison group students, I am often asked what element about the Academies' model accounts for this difference. The only honest answers I can give are either "the whole set of elements that comprise the model" or "I do not know." Of course, one can form judgments. I talk to many Academies' teachers and administrators, and I have my own opinions about what features of the model are most critical for success, but these are based on judgments, not hard evidence.

A fourth problem that must be anticipated is based on program operators often being defensive about having evaluations conducted. However objective and fair evaluations may be in their design, negative findings will almost surely negatively affect the operation of a program and often the program staff's jobs. Such negative findings will often lead to challenges about the evaluation's design and methodology; positive findings, in my experience, never raise such questions. In addition, once the objectives of the study are understood, especially the variables that will indicate participant success or failure, program operations may be adjusted to improve the likelihood of success. One such adjustment is in participant selection, that can result in the problem of "creaming." Better performing students will more often provide more positive results. Their selection also works directly against the objective of serving all students fairly and equally.

Still another problem of conducting such evaluations is the quality of data often available. High schools and school-to-work programs are not havens for carefully kept records and precise data. Often, it is difficult to obtain consistent and complete data on even the simplest of variables. Attendance may be kept five different ways within a city. Credit systems vary from one district to another. Test scores can be reported in many forms (e.g., scale, grade equivalent, local percentile, national percentile, and normal curve equivalent). Such problems often make comparisons between one setting and another difficult. Even within one district or school, it is surprising how incomplete records often are, particularly for students who disappear for some period of time. Few high schools can provide precise data on dropouts; if a student stops attending, and there is no transcript request from another high school, at some point it is simply assumed that the student dropped out. The problem of low-quality databases in this arena can overwhelm the best of designs.

2. Timing

One of the questions posed for this paper was how far the school-to-work initiative has to progress before a net impact study should be implemented. There is no pat answer to this question, but there are some useful guidelines. One rule of thumb is that examining any program's participant effects during its first year of operation is likely to lead to either inconclusive or incorrect findings. Any program has startup problems; and assuming what is sought is the potential for positive impact, this is likely to take two or three years to be realized. Thus, I would recommend waiting this long to examine student outcome measures.

This does not mean nothing should be done in the interim, however. As discussed above, an important element of examining participant impact is looking at prechanges and postchanges, whether or not comparison or control groups are employed. Predata are best obtained either during the "pre" period or shortly thereafter. In fact, if predata include interview or questionnaire data, they must be obtained before the program begins. Even that sought from records may be lost if not obtained early on. For example, many high schools and districts do not retain attendance data much beyond the school year to which they pertain, so if they are not obtained at that point, they may well be lost forever. While other school and program records are likely to be available for a long period, time often erodes these as well. Thus, a net impact evaluation should be designed and employed as soon as possible.

Such an evaluation should include measures other than participant outcomes, which also argues for early data collection. As discussed above, implementation information should be one important ingredient in such an evaluation, and this information needs to be obtained on an ongoing basis. The same is true of policy impact data. In fact, the evaluation should focus on these matters most strongly in its early phase and move gradually to a more central focus on participant impacts over a several-year period. The information that comes out of the implementation and policy impact portions of the evaluation is needed to help in the design of the participant effect study. Assuming a sampling system is used, it will be information from these earlier phases that will inform the choice of States and localities to be included in the indepth portions of the study.

The length of time over which participants should be followed is also an important issue. To address the effects of a school-to-work program, obviously it is necessary to collect data for the period of time the participants are in both of these phases. As discussed above, it is desirable to gather some information about participants before they enter the treatment, particularly on in-school measures available through records (attendance, credits, GPAs, and disciplinary referrals), so that comparisons and postcomparisons are possible. It is also important to gather data on postprogram outcomes, which in this case means through completion of postsecondary training and at least initial employment. While the length of the in-school treatment will vary with the program design, in most cases this implies the need to follow participants for a four- to five-year period and in some cases, for six or seven years.

3. Department of Education/Department of Labor Guidance and Incentives

Another question posed in the guidelines for this paper was what guidance, information, and incentives the Departments should offer the school-to-work community to facilitate such an effort. There seem to be several answers to this question. Presumably, some reporting requirements are already in place for States receiving funds through the Act. Likewise, presumably local sites receiving these funds must report to the States. These reports can include at least some of the broad kinds of information discussed above, especially regarding implementation and policy impacts. This should be a simple prerequisite of funding.

Once some of the details of the study are determined, information about these can be provided to States, with encouragement to pass this information on to localities so that both can begin to prepare the way for the indepth evaluation. Some States and localities will probably be more open to such evaluation than others, and this feedback can be useful in developing sampling strategies. The information that will be gained through the evaluation concerning the relative effectiveness and efficiency of alternative approaches may provide an incentive to States to become involved with the indepth portion of the study.

Beyond this, a participant effect study will require hiring a firm that is expert in such matters and providing it the necessary funding to conduct such a study. Such evaluation is generally expensive and difficult, and I know of no way to conduct it without such support.

V. Conclusions

The School-to-Work Opportunities Act launched an important body of reforms. It has charted a new course for the Nation with respect to preparing young people for work and careers. The breadth and depth of change it envisions and encourages is in a number of ways unprecedented in the Nation's history. These are important changes to the young people who will be directly affected and to the Nation's economy. Thus, an evaluation of the net impact of the Act is an important undertaking.

There are many issues to be considered in designing and implementing such an evaluation. While not all the details can be dealt with in a paper of this size, I have tried to summarize these issues here. I have reviewed the substantive content of the policy reforms as I see them, presenting a discussion of ten key elements of this reform. I have offered two broad hypotheses to be tested in such an evaluation and described the kinds of data to be collected in four categories—implementation, policy impacts, participant effects, and comparative efficiency-to test these hypotheses. I have provided suggestions for the instruments to be used for such data collection in each of these categories and recommended a sampling strategy. I have presented three participant effects evaluation design options and discussed the advantages and disadvantages of each. All have limitations. I have reluctantly recommended use of the comparison group design. I have discussed a variety of implementation problems, offered recommendations with respect to the timing of the study related to its various categories of data collection, and described the support the Departments can provide to facilitate this work.

One can summarize the central purposes of this evaluation as follows:

- To learn to what degree and with what quality the policy and program elements of the STWOA are being implemented in States and regions and localities throughout the United States, what the barriers to success have been, and how to address them

- To learn what the STWOA impacts have been on policies at the State and regional and local levels

- To learn what changes in participant performance can be associated with STWOA programs

- To assess the efficiency of policy in this arena and the conditions under which cost-benefit ratios can be maximized

This is an immensely important task. Given the dominance of the global marketplace in the world today and the critical importance of employee skills in any nation's competitiveness, it is no exaggeration to say that the STWOA has the potential to influence the course of the Nation. The issues discussed herein deserve long and judicious thought, and any decisions regarding them need to be arrived at through careful discussion. If this paper can help to promote such thought and discussion, it will have served its purpose.

Net Impact Evaluation of School-to-Work: Desirable but Feasible?

Robert W. Glover
Christopher T. King
Center for the Study of Human Resources
The University of Texas at Austin

I. Introduction and Background

Several months ago, we received a call from the vocational director of a medium-sized Texas school district. She had just made a presentation urging her superintendent to reorganize the curriculum of a high school around career pathways and to build a school-to-work system. She argued that these measures would improve school attendance, student achievement, school completion rates, entry rates into postsecondary learning, and job prospects for the district's graduates. Her superintendent responded with a simple question: "Where is the evidence?" After the meeting, she was desperately calling in pursuit of information to back up her claims.

Unfortunately, the answer to the superintendent's question is all but simple. Due to the relative newness of school-to-work approaches, the long-term nature of the treatment, the varied and evolving nature of school-to-work models, the diversity of implementation practices, and the lack of consistency in terminology, definitions, and other factors, a clear and conclusive body of evidence is not yet available regarding the effectiveness of school-to-work initiatives.

Ideally, one would like to be able to compare the outcomes for two groups of students, who differed only in whether they received a given school-to-work treatment. One of the most scientifically rigorous ways to accomplish this is to use large-scale field trials in which eligible students are randomly assigned either to a treatment or to a control group that does not receive the treatment. Such approaches have become commonplace in evaluating job training, welfare-to-work programs, and some other areas, but they remain relatively rare in education. However, large-scale field tests with random assignment have been used recently with some success in education, proving that they are feasible under the right circumstances. Examples include a Tennessee study of the effectiveness of small classes in raising student achievement (Tennessee State Department of Education, 1990), a field test of approaches to drug abuse prevention in California (Ellickson and Bell, 1990), and an ongoing study of high school career academies by Manpower Demonstration Research Corporation (MDRC). In addition, Crain, Heebner, and Si (1992) studied the effectiveness of Career Magnet Programs in New York City making use of "natural data" from the lottery system already in place for assignment of students to the magnets. Hayward, Tallmadge, and Leu (1992) were able to use an experimental design with random assignment to evaluate seven of ten dropout prevention demonstration projects (cited in Stern et al., 1995, p. 83).

Unfortunately, the standards of proof used in the education community are not always high. Some American educators seem to jump from one fad to the next, seeking the miracle cure for student performance problems. With support from several major foundations, special funding under the School-to-Work Opportunities Act (STWOA) of 1994, and devoted attention from President Bill Clinton and the Departments of Education and Labor, school-to-work has been a favored approach for the past few years. However, such fortunate circumstances may not last. If the school-to-work movement is to persist and to compete effectively with more established programs and more immediate needs in workforce development in a block grant environment, it must accumulate a base of reliable evidence to demonstrate its effectiveness.

While the Federal government has been a source. of initial funding and direction for school-to-work programs, the STWOA clearly envisioned turning program decisionmaking and funding over to states and localities. Thus, the primary audience for any evaluation findings over the long haul is likely to be the various state and local school-to-work partners.

This paper discusses issues and alternative approaches to building a body of evidence regarding the net impact of school-to-work initiatives, especially examining what role(s) field tests with random assignment and similar approaches may serve in this endeavor.

1. Existing Evaluations of School-to-Work Initiatives

A growing body of case studies and process evaluations of school-to-work programs is becoming available (Goldberger, 1992; Goldberger, 1993; Hamilton, Hamilton, and Wood, 1991; Hamilton and Hamilton, 1993; Hamilton and Hamilton, 1994; Hoerner et al., 1992; Pauly, Kopp, and Haimson, 1994; Goldberger, Kazis, and O'Flanagan, 1994; Mathematica Policy Research, 1994; Kopp and Kazis, 1995; Owens et al., 1995; Bailis, 1995; and the Academy for Educational Development, 1996). In addition, there is a growing literature on the implementation of Tech Prep programs since the passage of the Perkins amendments of 1990, which provided Federal funding to promote the implementation of Tech Prep programs (Bragg, Layton, and Hammons, 1994; Jackson, Dial, and Strauss, 1994; Silverberg and Hershey, 1995; Hershey, Silverberg, and Owens, 1995). Through such documents, a rich collection of examples, testimonials, anecdotes, lists of "promising practices," and other information regarding the implementation of school-to-work programs is being accumulated. To date, however, few of these evaluations have yet focused much attention on the outcomes of school-to-work initiatives.

Probably the best compilation of empirical evidence regarding the effectiveness of school-to-work efforts, broadly conceived, is found in Stern et al., School to Work Research on Programs in the United States (1995). The authors review available evidence regarding the effects on student performance in school and subsequent employment and earnings of several programs, including various types of secondary vocational education, cooperative education at the secondary and postsecondary levels, career academies, and even of nonschool supervised work experience. They also review known evidence regarding effectiveness of program components such as mentoring and the integration of academic and vocational education. On the whole, their findings on student gains in achievement and persistence in school are positive, whereas the effects on subsequent earnings and employment are mixed.

An example can be found in the evaluations of career academies. Stern, Raby, and Dayton summarize the results of available evaluations of career academies in California, Philadelphia, and New York (1992, pp. 56-71). Most of the evaluations used quasi-experimental designs with comparison groups, mostly drawn from the same high schools. Generally, the comparisons revealed that students enrolled in the academies recorded better attendance, failed fewer courses, earned more credits, obtained better grades, and enjoyed higher graduation rates than did students in comparison groups. In most cohorts, academy students were half as likely as their comparisons to drop out before completing high school. Results of followup studies showed mixed results with respect to participation in postsecondary schooling. One evaluation found that academy graduates were much more likely to be enrolled in postsecondary education; one found them less likely to be enrolled; and a third found no difference (pp. 66-69). Likewise, data on differences in hourly wages within a year or two from graduation were mixed and generally not statistically significant, but academy graduates worked 3 hours per week more than the comparison groups (p. 63).

Stern, Raby, and Dayton carefully noted the limitations of the evaluation results that they called "suggestive and promising." Because no career academy had been evaluated experimentally, it is possible that unknown selection effects may have influenced these findings. Further, these evaluations did not test which of the various elements of the academy program—school-within-a-school structure, integrated curriculum, career focus, related employment, or mentors-make the most difference (pp. 57 and 71). Nevertheless, these evaluations played a major role in securing funding from the California legislature for expansions of the state's career academy pilot program.

In California, initial capital costs of each academy were in the $50,000 to $75,000 range. Ongoing annual incremental operating costs were estimated at $750 to $1,000 per student, or 20 percent above California's average cost for a high school student (Stern, Raby, and Dayton, 1992, pp. 78-79). These costs include expenditures to provide teachers common planning time, modest reductions in class sizes, coordination with business partners, instructional aides, staff development, transportation, and maintenance of equipment. Even so, researchers estimated that the programs' benefits in terms of additional expected lifetime earnings as a result of higher high school graduation rates outweighed the costs (Stern, Raby, and Dayton, 1992, p. 61).

Currently, three major national school-to-work evaluations are under way. Mathematica is conducting a multiyear study of Tech Prep programs, as well as an evaluation of school-to-work partnerships. The third is a national evaluation of career academies conducted by MDRC.

Tech Prep Evaluation

In October 1992, **Mathematica** began a five-year evaluation of the Tech Prep Education Program for the Department of Education, focusing on two objectives. First, the study tracks the development of Tech Prep programs nationwide in terms of the number, characteristics, and activities of programs and the populations they serve. Second, in case studies of selected sites, the study is measuring the progress of Tech Prep students and identifying effective practices.

School-to-Work Implementation Evaluation

In September 1995, **Mathematica** started a five-year evaluation sponsored by the Departments of Labor and Education and the National School-to-Work Office to examine the implementation of school-to-work partnerships in seven to nine states receiving implementation grants. As part of the study, eighty-five students in each of thirty-two randomly selected partnerships across the states studied will complete a questionnaire at the end of high school, and then they will be interviewed about eighteen months after graduation. Information will be collected on students' access to and participation in school-based and workplace activities in the school-to-work system, and on changes in secondary, postsecondary, and employment outcomes achieved by successive cohorts over time. Information will be collected on cohorts of high school seniors in the years 1996, 1998, and 2000 to ascertain changes.

Career Academy Evaluation

MDRC's evaluation of career academies may be the only current school-to-work evaluation that uses a net impact research design with random assignment. Begun in 1992, this project is a demonstration and evaluation of ten career academies across the country. For the demonstration, participating academies have upgraded their recruitment efforts to attract approximately double the number of eligible students the academies had the capacity to serve. To accomplish the practical requirements of the study, MDRC sought sites that would help maximize the number of students in the research sample. Only sites with a capacity to recruit 100 to 110 students and to enroll at least 50 to 60 students each year were eligible. Since at-risk youth were of special concern, the study focused only on career academies that included students at risk of failing in a regular high school setting. MDRC choose school sites in which there were few alternative academies or academy-like programs available that would minimize the contrast between academy and nonacademy environments (and thereby tend to understate the impacts from participating in the academy).

MDRC selected sites that contained the defining structural features of the career academy approach, which the researchers identified as a school-within-a-school organization, integration of academic and vocational curricula, and partnerships with local employers to provide support for school-based and work-based learning activities. Within this framework, there was considerable diversity of practices among the academies studied. The career academies selected were not necessarily the best sites, but rather ones that offered a credible test of the career academy approach.

To qualify as candidates for the study, the academies selected had to enroll students no later than tenth grade. A school-within-a-school approach was defined as block scheduling a team of at least four teachers with a group of students in at least three classes per semester through junior year and at least two classes in the senior year. The schools made shared planning time available for participating teachers.

The MDRC study is being supported by a consortium of sponsors including the Labor and Education Departments and thirteen private foundations. To date, the $7.5 million committed to the project is sufficient to study the students through high school. Additional funding is being sought to collect followup information on employment and earnings and participation in postsecondary learning for three years beyond high school. Existing studies of career academies have shown mixed or insignificant postschooling results (Stern, Raby, and Dayton, 1992).

2. School-to-Work Evaluation Efforts by States

While it is premature to offer a comprehensive assessment of the states' evaluation efforts, a quick review of a handful of selected state plans for evaluating their school-to-work programs, gleaned from their grant proposals, suggests several insights. First, although some states have given considerable thought to their information needs and the types of approaches that might be required, for the most part, their evaluation plans are quite brief (e.g., three to five pages) and appear to be drafted in response to specific language contained in the request-for-proposals. Second, these plans have a number of common features, including reliance on an external evaluator for major evaluation tasks and development of integrated information systems and performance indicators or benchmarks. None of the states seems to have an existing evaluation framework into which their school-to-work efforts would be incorporated; rather, they are having to construct data systems and evaluation capacity largely de novo. In several plans, feasibility studies (e.g., for using unemployment insurance wage records to document postprogram labor market experiences) were envisioned as the necessary first step. Third, STWOA and National School-to-Work Office proscriptions notwithstanding, the states apparently enjoy wide discretion in terms of data elements to be captured, definitions of key terms, time periods (e.g., for followup), and the specific measures to be tracked, among others. Wide variation is likely to characterize states' evaluation approaches and findings, as well as their programs, if these plans are any indication. Finally, of the school-to-work evaluation plans examined, only New York envisioned conducting a net impact tracking study of some of its efforts in some of its school districts (including the "Big Five Cities"), complete with random assignment to treatment and control groups. Left to their own devices, it is clear that few of the states would be pursuing experimental (or even quasi-experimental) evaluations of their school-to-work initiatives.

II. Conceptual and Policy Issues

1. Basic School-to-Work Models and Typologies

To date, most of the energies of the school-to-work movement have been focused at the high school level, as most of the models described in this section reflect. This focus is understandable for a movement that was stimulated by concern for those who were not continuing their education beyond high school. Also, since the STWOA was signed into law merely two years ago, many of the participants are still high school students.

Pauly, Kopp, and Haimson have classified school-to-work programs into five basic models: (1) career academies, (2) occupational-academicclusters, (3) restructuredvocationaleducationand co-operative education programs, (4) Tech Prep programs, and (5) youth apprenticeships (1994, pp. 9-10). Descriptions of these models follow:

- ***Career*** Academies-They are schools within a school that group together a team of teachers with high school students in a three- or four-year program integrating academic learning with the study of an industry and the careers of the people who work in it (such as electronics, health care, or financial services). Local employers provide classroom speakers, industry tours and job shadowing opportunities, mentors and summer internships to introduce students to the industry. The career academy model is designed to provide a personalized learning environment in a school-within-a-school setting in large urban high schools.

- ***Occupational-Academic*** Clusters-These clusters organize high schools into several schools within schools, each focused on a particular theme or industry. These clusters can be large-scale efforts to offer all students in a high school a choice among several different pathways, each one based on a sequence of related courses tied to a cluster of occupations. Students are usually exposed to a wide variety of careers before choosing an occupational cluster and they may switch clusters during the program. Each cluster offers occupation-related courses in which students receive training in broad, work-related skills after taking introductory career exploration courses. Academic and occupational instruction are integrated and applied learning techniques are sometimes used. Work-based experiences enable students to explore potential careers.

- ***Restructured VocationalEducation and Co-OperativeEducation***—This approach reshapes traditional vocational education programs by providing earlier and broader opportunities to learn about careers through job shadowing and visits to workplaces, structured reflection on students' workplace experiences, and closer linkages between students' occupational and academic courses. The programs aim to include a larger and more diverse group of students in vocational programs, making career exploration a central part of their education and preparing them for a wide range of career opportunities rather than narrow occupational specialties.

- ***Tech Prep Programs***—These programs upgrade the general track and vocational high school curricula to emphasize technology-related instruction in science, mathematics, and other courses. Coursework includes hands-on application ofworkplace problems. Tech Prep aligns high school courses with requirements for postsecondary technical occupational training programs at community colleges and universities. High school students can receive college credit toward an associate's degree based on courses they take in high school through arrangements for dual enrollment, credit-in-escrow, or credit-by-examination procedures. Work-based learning is not usually included as a part of Tech Prep programs.

- ***Youth*** Apprenticeships-These programs use the workplace as a learning environment to provide students with competencies in technical skills and related math, science, communication, and problem-solving skills. Students learn by doing in paid employment in a structured training plan with an expert adult mentor, coach, or supervisor who works closely with them on job-related skills and general employment-related skills. Classroom vocational instruction and related courses that integrate academic and vocational learning are part of youth apprenticeships, and most programs link secondary and postsecondary institutions to provide such instruction. Qualified completing students receive a recognized occupational credential.

The list of program types above is neither complete nor exhaustive; rather, it offers a sample of some of the major current efforts. Additional classifications include:

- ***School-Based*** Enterprises-They produce goods or services for sale or use by people other than the participating students (Stern et al, 1995, p. 36). This model is attractive to schools in rural areas without a strong base of employers to provide work-based learning.

- ***Career Magnet Programs and*** Schools-They were established to encourage voluntary school integration by providing an attractive program or school geared to career preparation in an occupational cluster or area of focus, usually offered in collaboration with industry partners.

- ***School-to-Apprenticeship Programs***—These programs place high school seniors into apprenticeship programs sponsored by employers and/or employers and a union. The apprenticeships are registered with the Department of Labor or state apprenticeship agencies. The program gives high school students a jump start on apprenticeship (U.S. Congress, Office of Technology Assessment, 1995, p. 59).

- ***Clinical Training***—Such training has been most commonly used in medically related occupations. Students in the program move through a series of worksite positions that provide training and unpaid work experience, generally under the supervision of school personnel. Usually, participation in a course of study, work experience, and passing scores on an external examination are required to obtain a license to work in the occupation (U.S. Congress, Office of Technology Assessment, 1995, p. 58).

Overall, these nine models can be sorted into two basic typologies of school-to-work programs, depending on their primary emphasis: school reform or occupational preparation. Of course, these categories are neither contradictory nor mutually exclusive; high-quality schools are surely a foundation element of any workforce development system. Key indicators distinguishing these two typologies are the roles for work-based learning and whether sufficient preparation is offered in occupational or job-specific skills to secure employment. Youth apprenticeships, school-to-apprenticeship programs, and the clinical training models incorporate the development of occupational skills in work-based learning. Most of the other models have used worksite components to provide career exploration opportunities, to foster generic workplace skills, or as incentives to motivate high school students to work hard and achieve in school.

The models described above offer ideal or theoretical types. In practice, there is considerable diversity among programs of the same model. MRDC researchers reported that while most programs generally contained all the key elements associated with the model they represent, many programs included additional features associated with the other models, resulting in a wide variety of hybrid programs (Pauly, Kopp, and Haimson, 1994, p. 17). Thus, regardless of the model, the specified components are not necessarily tightly defined within models, much less between them-a factor that complicates any evaluation effort. For example, although the Tech Prep model does not generally include worksite learning, a few Tech Prep programs have established work-based learning components. Similarly, youth apprenticeships have been added as opportunities for students in some career academies.

The crossover of elements or hybridization among the various models has likely increased substantially since passage of the STWOA as states and localities have competed for school-to-work funding. The STWOA elaborated on a list of components to be included in school-to-work programs supported by the Act. Nearly two dozen components were organized under the headings of school-based learning, work-based learning, and connecting activities. Not only do so many components introduce too much variation to make a comprehensive evaluation financially feasible, but it is questionable that we could find even a single site with all components fully in place. Thus, we need to determine which among the nearly two dozen components specified in the STWOA are most essential. In sum, what are the defining characteristics of school-to-work?

There is no consensus on which elements are most important. For example, program staff in the 16 sites studied by MDRC identified the most crucial elements of their programs as integration of academic and occupational learning in school, applications-based instructional methods, a strong and demanding instructional program, extra support for students through clusters or a school-within-a-school approach, and high-quality workplace learning experiences for students (Pauly, Kopp, and Haimson, 1994, p. 176).

Stern et al. emphasized the following three components as most important: integration of school-based and work-based learning, combined academic and vocational curriculum, and linking of secondary and postsecondary education (1995, p. 12).

Dayton, in his paper for this volume, provides yet another articulate list of ten "principles and elements" of school-to-work that includes the following:

- Collaboration
- Pedagogy
- Service to all students
- Technology and the high performance workplace
- Teaching of all aspects of the industry
- Standards, assessment, and credentialing
- Fostering of systemic change in schools
- Employer involvement
- Establishment of work- and community-based learning
- Professional development, resources, and technical assistance

It is clear that experts disagree. Thus, one of the most important conceptual issues to address is quite basic: what is school-to-work?

2. Judging the Relative Effectiveness of School-to-Work Approaches

There are far too many program models and components to be feasibly accommodated in any evaluation undertaken at reasonable cost. Choices must be made. Ideally, a process of consensus should be developed to select the most promising school-to-work models and the most critical outcomes to use in an evaluation. In the absence of such a process, we raise and discuss two selections for consideration: youth apprenticeship programs that have stressed the occupational training and workforce development aspects of school-to-work, and occupational-academic cluster programs that give greater emphasis to high school reform processes. It is not at all clear that net impact analysis using random assignment techniques is the most appropriate evaluation tool for either model.

Youth Apprenticeship

The youth apprenticeship model provides preparation for full entry-level qualifications in an industry through a coordinated program of school-based and work-based learning featuring career counseling, integrated academic and occupational instruction, structured training and mentoring at the workplace, progressively higher skill advancement and pay, and the opportunity to earn an industry-recognized skill credential.

Youth apprenticeship was the flagship program of the school-to-work movement in the early 1990's. It was designed to provide high-standard occupational skill development for high school students not planning to attend college. As it turned out, significant percentages of the participating students subsequently decided to enroll in college, causing some of the sponsoring employers to question the investments they had made in training them.

Several states, notably Arkansas, Maine, Pennsylvania, and Wisconsin, have invested in the approach, and a few dozen localities have established youth apprenticeship programs, some with demonstration funding from the Department of Labor. Students normally entered these programs at the start of their junior year and were expected to continue in the training for at least a year beyond high school graduation if the occupation called for it. Considerable efforts have been made to recruit employers, structure the worksite learning and coordinate it with school-based learning, provide mentors, and adjust student schedules to accommodate work-based learning during the school year.

The preliminary conclusions on this initiative to date are that while youth apprenticeships offer excellent opportunities for some students and certain firms, they will be difficult to bring to scale to serve many students, due to program costs that are substantially higher than current education spending and due to the challenges of recruiting employer sponsors (U.S. Congress, Office of Technology Assessment, 1995; Orr, 1996). Regarding the outlook for expanding apprenticeships among employers who have not traditionally hired youth, most observers foresee little prospect of preferred employers offering significant occupational training to adolescent youth on any major scale any time soon. A recent report on work-based learning by the Office of Technology Assessment concluded as follows:

> Because the development of occupational skills requires greater effort on the part of employers, the employers are more likely to make the investment for students who are relatively mature, sure of their direction, and close to entering the labor market. Such students provide employers with a better chance of recouping their training costs (U.S. Congress, Office of Technology Assessment, 1995, p. 5).

Of course, most American high school juniors and seniors already work, but the jobs they currently hold are unconnected to school and generally offer little prospect for learning or advancement. This information has led to an alternative strategy-to develop work-based learning by upgrading the learning in jobs where youth are already employed, for example, by building in the development of competencies and foundation skills and personal qualities outlined by the Secretary's Commission on Achieving Necessary Skills (SCANS) in 1991 and other strategies. Examples of this approach to upgrade youth jobs include the WorkPlus Project, administered by Public/Private Ventures (MacAllum and Ma, 1995), and the Consumer Service Management National Youth Apprenticeship Demonstration Project, sponsored by McDonald's with other corporate partners in collaboration with Northern Illinois University.

In any case, the future status of employer participation in youth apprenticeships will depend heavily on employers' assessments of their own benefits and costs of providing the training positions (Mathematica, 1994, pp. xi-xii). A net impact analysis focused on outcomes for students would not address such employer concerns. However, the National Employers Leadership Council and ten states reportedly are conducting return-on-investment analysis of firms participating in school-to-work transition programs (National Governors Association, 1995, Table 1).

Youth apprenticeships remain in an experimental stage. Researchers at Jobs for the Future, one of the leading national organizations advocating for youth apprenticeship, recently proposed significant modifications of the model. They recommended that students begin the work-based portion of their youth apprenticeships in their senior year of high school, after completing core requirements for graduation, and that postsecondary institutions become the primary providers of the technical classroom instruction for the apprenticeships. These changes are needed, they conclude, to "make it more consonant with the realities of both high school funding [and scheduling] and young people's career decisionmaking (Goldberger and Kazis, 1995, p. 21). In short, the youth apprenticeship model may be evolving in significant ways that make it an unattractive candidate for a costly long-term net impact analysis at this time.

Occupational-Academic Clusters

Among existing school reform program models, the most appealing candidate for investigation may be occupational-academic cluster programs. There are several reasons for focusing on occupational-academic clusters. Occupational-academic clusters have been less researched than Tech Prep programs or career academies. Cluster programs are likely to expand, reaching potentially large numbers of students. The provision for career majors to begin not later than eleventh grade as specified in the STWOA [Section 4 (5)] has given impetus to the formation of occupational-academic clusters. According to an analysis published by the National Governors Association in July 1995, thirty states were establishing uniform statewide career clusters, which will facilitate replication of occupational-academic cluster programs (National Governors Association, July 29, 1995, Table 5). Like career academies, occupational-academic clusters offer a means to restructure large urban high schools into smaller, more personal, and supportive learning environments to encourage student engagement in school and persistence to graduation. Clusters can facilitate the integration of academic and occupational learning (Grubb et al., 1991). Moreover, several occupational-academic cluster programs are already established and may be suitable sites for a net impact study. Finally, the approach to restructuring high schools in occupational-academic clusters is consistent with the research findings and recommendations made by the National Assessment of Vocational Education (Boesel and McFarland, 1994), Jobs for the Future (Goldberger and Kazis, 1995), the National Center for Research in Vocational Education (Grubb, 1995; Stern et al., 1995), and the National Association of Secondary School Principals (1996).

The ideal high school clusters program aims at high academic standards and expectations, with each broad cluster encompassing preparation for some occupations that require postsecondary education. Courses satisfy high school graduation requirements and selected courses are accepted for college credit by postsecondary institutions, through written articulation agreements. Occupational-academic cluster programs provide students the choice of several schools within schools offering personal and supportive learning environments. They integrate academic with vocational applications within broad occupational themes. They include counseling to assist students with their career decisions, usually including a course or seminar focused on career development prior to entering the cluster that introduces students to the various available clusters. Work-based learning in various forms is offered to provide opportunities for career exploration and to motivate student performance in school.

Just as with all of the school-to-work models, occupational-academic cluster programs vary significantly in practice and are customized to fit state requirements and local circumstances. What is needed to begin an evaluation is agreement on a common framework specifying the essential components of occupational-academic cluster programs. Substantial care would be required to assure a reasonably standardized approach across the sites.

3. What Outcomes Should Be Expected From Various School-to-Work Models and Treatments?

The key hypothesis for any net impact evaluation of any school-to-work initiative is to what extent is it producing the desired outcomes for youth?

Some outcome measures are more appropriate for certain versions of school-to-work than others. For example, improvement in employment and earnings is unquestionably an appropriate outcome indicator for school-to-work approaches that emphasize workforce development, such as youth apprenticeships. However, there is some controversy about whether initial gains in employment and earnings are suitable outcome measures for school reform approaches such as occupational-academic clusters. High school clusters often avoid teaching job-specific or occupation-specific skills, deferring such training to postsecondary levels. In view of this information, it may be inappropriate to expect net gains in employment and earnings immediately after high school (Grubb et al., 1991, p. 83).

Others contend that emphasis on broad industry-related clusters does not preclude the possibility of students gaining specific marketable occupational skills. Jobs for the Future researchers Goldberger and Kazis argue that even when the emphasis in high school is placed on broad intellectual preparation rather than occupation-specific training, participating students can acquire specific marketable skills, credentials, and employer contacts and references through their work-based learning experiences that provide them advantages in the labor market after high school. Although the researchers present no systematic data to prove their point, they illustrate it with examples from the ten projects they monitored (Goldberger and Kazis, 1995, p. 18). Additionally, they argue that the explicit skill standards and credentialing procedures under development by industry groups can help drive the design of programs organized around career themes and facilitate the accreditation of the specialized competencies that students acquire in career-related programs of study (Goldberger and Kazis, 1995, p. 34).

Unless the school-to-work movement is badly misnamed, improved employment and earnings for participating graduates must be considered as a primary outcome for most, if not all, school-to-work models. While it may not be appropriate to expect gains in employment and earnings immediately after high school graduation, gains should be anticipated to materialize at some point following postsecondary learning. This factor calls for a net impact evaluation focused more on longer-term labor market outcomes.

Just as we must decide what constitutes school-to-work-broad-based school reform or selected essential elements-we must also determine which of the many possible outcomes are to serve as appropriate evaluation measures.

4. School-to-Work: An Evolving and Expanding Vision

> Visioning is a continuous process. As you move through the various stages of the system-building process, this vision will reflect the lessons learned (National School-to-Work Office, The School-to-Work Template: A Guide to Building School-to-Work Systems, not dated).

The school-to-work movement has become remarkably dynamic over the past decade and shows no sign of stabilizing any time soon. The aims and claims of school-to-work advocates have expanded considerably, to a large extent shifting away from the movement's early focus and structure.

In the 1980's and early 1990's, school-to-work was advocated as a means to improve the employment preparation and prospects of the "noncollege bound," the "forgotten half" or the "neglected majority" (e.g., Commission on Skills of the American Workforce, 1990; William T. Grant Foundation, 1988; Parnell, 1985). Reports called attention to the differential in public investments between college-bound and noncollege bound youth (William T. Grant Foundation, 1988; Commission on Skills of the American Workforce, 1990; U.S. General Accounting Office, 1991). Special focus was given to subbaccalaureate labor markets. The youth apprenticeship model received primary attention, spurred by a series of grants made by the Department of Labor under the School-to-Work Transition/Youth Apprenticeship Demonstration. This demonstration began in September 1990 with grants to six organizations. In fall 1992, funding for five of the original programs was renewed, and ten additional organizations were provided two-year grants.

In its more recent incarnations, school-to-career systems are being promoted as part of broader school reform measures designed to serve *all* students, regardless of their college attendance plans (Goldberger and Kazis, 1995). Although those not planning to enroll in college may benefit most from participation, school-to-career initiatives are now viewed as appropriate and valuable for all students.

In part, the evolving vision of school-to-career systems has been a response to parents who resist any reform that sounds as though it may preclude college attendance for their children and its associated access to skilled jobs offering higher wages. The new vision aims to avoid tracking, which limits students' options and risks stigmatizing the initiative. American school-to-work advocates who were inspired by European apprenticeship systems in Germany and Denmark have noted that the Europeans are changing to build college options into their own apprenticeship systems. In part, the new vision is a reaction to research findings of limited payoffs from past secondary vocational education in the United States (Stern et al., 1995). Finally, the vision is a response to the small size and slow growth of youth apprenticeships. A recent report by the Office of Technology Assessment surveyed employer involvement in fifteen school-to-work transition programs (of which ten were youth apprenticeship programs), During the previous two years, the median growth rate of participating employers was only six employers per year (U.S. Congress, Office of Technology Assessment, 1995, p. 76).

Currently, this evolving school-to-work vision includes several elements, which emphasize the following:

- Serving all students, whether or not they plan to attend college.

- Building a system, not a program. In part, developing a system involves building on the best elements of existing programs such as cooperative education and others and connecting them in a coherent system, along with newer program efforts such as youth apprenticeship programs, Tech Prep, and academic-occupational cluster schools. It means aligning with broader school reform efforts, including Goals 2000 initiatives, the academic standards movements coming from professional groups in various academicdisciplines, state efforts to revise and update curriculum frameworks and outcomes, and other efforts. A school-to-careers system influences teacher preparation to produce teachers with the knowledge and skills to teach in new ways and to work effectively with heterogeneous classes of students. It requires revising school schedules to facilitate thematic learning, integration across academic and technical subjects and to allow common planning time for teachers and work-based learning for students. The school-to-careers system brings together school reform, economic development and workforce development, creating a "seamless system" that serves both youth and adults with a common credentialing system based on industry skill standards. Building a comprehensive system involves establishing linkages with youth development agencies and youth-serving agencies, with alternative learning centers serving youth who have dropped out of school, and with agencies serving disabled youth-all of whom can provide support services to help youth at risk to succeed. It means implementing policies and frameworks at the state level to support and foster school-to-work.

- Eliminating all tracking, eventually including separate vocational and academic tracks, and replacing them with thematic high school programs of study that emphasize practical education for all. Berryman and Bailey (1992) have characterized the traditional educational system as "decontextualized academics and academically debased vocational education" (p. 106). Advocates of this perspective perceive integrating vocational and academiceducation as more than just upgrading vocational education; it offers a means to reshape the academic curriculum as well, providing students with opportunities to apply what they learn. The process has the potential to shape a very different kind of high school-eliminating the "shopping mall high school" in which students shop around picking courses that strike their fancy and accumulating an incoherent program. It offers potential to improve the teaching of all subjects, to enhance the engagement of students, and to reduce the isolation of teachers and students. It provides specific roles for employer participation—reinforcing the importance of learning both technical and academic competencies by offering the extrinsic rewards of summer internships and future employment and by testifying that what is learned in school is useful in other settings (Grubb et al., 1991, pp. 86-91).

- Integrating work-based learning into the core curriculum for high school students primarily to allow them to explore careers and to develop generic work skills. Under this approach, the development of occupational skills generally is deferred to the postsecondary level (U.S. Congress, Office of Technology Assessment, 1995, p. 15).

- Using instructional strategies that incorporate principles of active, experiential and applied learning, cooperative learning, and learning in context. These changes imply revisions in assessment practices as well as in curricula and instructional approaches and professional development to help educators to implement these approaches effectively.

- Leveraging funds and redirecting existing resources to implement the vision. Clearly, the ambitious vision of the school-to-career system will not be implemented on the basis of earmarked Federal seed monies alone. Total national funding for school-to-work, including Tech Prep, amounted to only about $30 per student spread across the nearly 12 million secondary-level students in the United States in FY '96. However, leveraging or redirecting other funds often requires considerable time and energy.

Clearly no state or locality has completely reached these ideals. However, the Federal government, resisting the temptation to impose any uniform approach on states and localities, has permitted and encouraged considerable flexibility and diversity. In effect, the components of school-to-work specified in the STWOA have been guidelines rather than mandates for change.

5. The Evolution of Programs Into Systems: Cautions for Evaluation

The expansion of focus from school-to-work programs to school-to-career systems has complicated the task for evaluation. Systems are more comprehensive and complex than programs. Systems fit into institutional contexts that are often nonreplicable. This feature may make systems far more idiosyncratic and thus less comparable across sites.

A systems approach may offer the school-to-work movement the most effective route for creating a lasting influence on schools. However, if comprehensive reform of the public education system becomes the primary goal of school-to-work rather than a slate of program goals and objectives, it is not at all clear that net impact evaluation based on random assignment of students is the appropriate evaluation vehicle.

III. Design Issues

One of the more important design issues was largely presupposed when the Department of Labor commissioned papers on alternatives for conducting net impact evaluations of school-to-work programs. Opting for an experimentally based evaluation of these programs precludes other options such as examining gross outcomes and performing relative program effectiveness analysis. The choice to pursue an experimental approach is widely endorsed in the evaluation literature (e.g., Betsey et al., 1985; Burtless, 1995; Fraker and Maynard, 1987; Stromsdorfer et al., 1985). However, the literature does not confer blanket support for random assignment field studies in all circumstances or for all types of interventions. In addition, some experts have recently emphasized using alternative impact estimation methodologies (e.g., Bell et al., 1995; Bowman, 1992; Heckman and Smith, 1995). Heckman's (1996) paper in this series even suggests that, if selection bias is more appropriately defined and if comparison group matching stresses preprogram unemployment rather than earnings, quasi-experimental approaches may do as well at estimating postprogram impacts as the more intrusive experimental ones. However, others (e.g., Blalock, 1990) have questioned the applicability of such quasi-experimental methods to youth populations and programs and, to the extent that their youth minimizes reliable information about their preprogram unemployment patterns, even Heckman's approach might not be feasible in the school-to-work context.

The relevant language in STWOA is fairly open in terms of approach and methodology. Section 402(b) states:

> Not later than September 30, 1998, the Secretaries [of Education and Labor] shall complete a national evaluation of School-to-Work Opportunities programs funded under this Act by grants, contracts, or otherwise, that will track and assess the progress of implementation of state and local programs and their effectiveness based on measures such as those measures described in subsection (a).

The measures described in section 402(a) hold few surprises. They do offer added detail on the key actors and target populations for whom progress must be tracked and for whom impact estimates might be estimated. For example, 402(a)(S) indicates that its authors were especially concerned about outcomes for "participating students and school dropouts, by gender, race, ethnicity, socioeconomic background, limited-English proficiency, and disability of the participants, and whether the participants are academically talented students . . . ." The outcome-based information required for these groups includes the following:

> (A) academic learning gains;
>
> (B) staying in school and attaining-
>   (1) a high school diploma, or a general equivalency diploma, or an alternative diploma or certificate for those students with disabilities for whom such alternative diploma or certificate is appropriate;
>   (ii) a skill certificate; and
>   (iii) a postsecondary degree;
>
> (C) attainment of strong experience in and understanding of all aspects of the industry the students are preparing to enter;
>
> (D) placement and retention in further education and training, particularly in the career major of the student; and

(E) job placement, retention, and earnings, particularly in the career major of the student.

This section examines selected design issues for evaluating school-to-work programs using an experimental approach with random assignment, including: the variation and stability of the program treatment to be evaluated, the optimal randomization point, outcome measures, target populations and subgroups, site-selection issues, sample size and representativeness, and timeframes. It does so largely from a practical vantage point, drawing on lessons learned in designing and conducting evaluations of related efforts. It stresses the special features of school-to-work programs as they now operate and as they might be expected to operate in the future.

1. School-to-Work Treatments or Models

The first issue addressed concerns which treatment interventions or models should be the focus of any such evaluation. There are several key aspects to this issue: treatment or model variation; treatment maturity or stability over time; and effects on the optimal point for random assignment.

Treatment/Model Variation

As indicated in Section II, there are at least nine major school-to-work program models-including career academies, occupational-academic cluster programs, restructured vocational education and cooperative education programs, Tech Prep, youth apprenticeship, and school-based enterprises-as well as an unknown number of hybrid programs that combine the main features of the known models. Moreover, these program models tend to have widely varying configurations of the two dozen STWOA program components, some of which may be more essential than others.

The school-to-work movement might be viewed as moving toward a continuum of program types, on one end of which are models that emphasize public school reform over work-related outcomes. At the other end are models emphasizing occupational preparation and transition to work, the school-to-work movement's original focus. Work-oriented school-to-work programs lend themselves more readily to experimentally based evaluations in which students are randomly assigned to treatment and control groups, while those aiming for major changes in public schools-including eliminating tracking, improving their curriculum and assessment practices, professional development of teachers, developing relationships with the employer community and other similar results-may not. To the extent that the latter type of initiative lends itself to random assignment, it is more likely to be the schools that are subject to random assignment, not the students. Whatever policymakers' preferences, an experimental design is not likely to be very helpful in gauging the efficacy of school-to-work in promoting broad-based public school reform.

Thus, school-to-work presents an environment within which there is substantial between- and within-intervention variation in what constitutes a given treatment or model-not to mention the high variability across states and localities. This variation creates numerous barriers to designing and conducting a useful impact evaluation. It also reinforces the need for conducting detailed process evaluations before sites are designated or particular school-to-work interventions/models are selected for inclusion in an impact evaluation. In fact, the Burtless and Moffitt papers in this volume suggest that process, implementation, or "overarching observational" studies should be conducted to supplement, as well as to complement, an impact evaluation.

Treatment Maturity/Stability

If school-to-work models or treatments are still evolving and undergoing rapid change, it will be difficult to evaluate them in any meaningful way. Only mature or well-established school-to-work models or treatments should be the focus of costly impact evaluations. Generally, programs that are less than three years old should not be considered viable candidates for inclusion in a net impact evaluation. In MDRC's Career Academies Evaluation, which began collecting information on the 1993-94 entering classes (cohorts), five of the ten academies studied were begun during the 1980's, four during the 1991-92 school year, and one in 1993-94 (Pauly and Thompson, 1993). On the other hand, some researchers have noted that becoming established may take far longer. Grubb et al., on the basis of their study of several schools that had implemented academic and vocational integration, concluded that "meaningful reforms will take, at a minimum, five to ten years of planning, implementation, and revision" (1991, p. 76). Likewise, a study of clinical training programs concluded that these programs went through five or more years of adjustments before achieving excellence (Bragg and Hamm, 1995).

If school-to-work has not yet settled into a stable operating mode, then evaluation findings are unlikely to be well received--or seen as valid-by either policymakers or the general public years down the road. The national Job Training Partnership Act (JTPA) evaluation is a case in point: by the time the JTPA evaluation results first became available to national policymakers in early 1993 (Bloom et al., 1993), certain aspects of JTPA policy (e.g., dropping cost standards and shifting to exclusive reliance on postprogram standards for adults) had changed substantially, and the program had undergone two major (1988, 1992) and one minor (1986) legislative reauthorization. After all, the JTPA programs being evaluated were the ones operating in the various sites as of 1986-87, not the program as it existed in 1993. Similarly, the ongoing national evaluation of the Federal and state Job Opportunities and Basic Skills training programs, conducted by MDRC and others, began several years ago and is likely to yield impact findings for policymakers only after the program has been eliminated or has been reformed beyond recognition.

With school-to-work programs, the variation is far greater and the likelihood of its models and treatments becoming stable any time soon is quite small. In addition, many of these school-to-work interventions, by design, are expected to last for several years, which leaves program staff plenty of time to make any number of very appropriate midcourse program adjustments.

On a basic level, policymakers may simply want to know the impact on certain schooling and labor market outcomes of participating in any of the school-to-work programs compared with not participating in a structured effort. Alternatively, they may primarily be interested in the net impact of participating in one school-to-work model compared to participating in another. Based on recent state and local program implementation, policymakers may only desire to obtain solid net impact estimates of participation in some of the school-to-work models, say Tech Prep or school-based enterprise. A variation on this example would be for policymakers to identify the models they feel are likely to be the more important ones in the future (those models thought to have staying power) and to concentrate their evaluation resources exclusively on these models. Yet another approach would be first to identify which of the essential components are dominant (most important) in practice, and then to evaluate varying school-to-workprogram streams featuring combinations of these interventions, similar to the JTPA study design (Bloom et al., 1988). Each of these intervention options has implications for such related issues as sample design and size, and the appropriate point for randomization, among others. In general, the more information is desired (e.g., impacts for more interventions, groups or impacts for one intervention versus another), the more complex (and more fragile) the design, and the larger the overall sample size required.

We can draw on insightful work by MDRC in analyzing the potential for using random assignment to evaluate Chapter 1 education programs and in its current evaluation which seeks to estimate the net impacts from Career Academy participation (Pauly and Thompson, 1993, pp. 14-23; Kemple and Rock, 1996). One of the ten important lessons that they learned during the course of designing and implementing their evaluations was that field studies with random assignment should be used "to learn about *concrete, well-specified* educational interventions, not broad system-wide reforms" (emphasis added). A related lesson was that "[c]aution should be used in developing random assignment field tests of educational interventions that use substantially different organizational structures or educational methods in different schools, provide substantially different amounts of services to students in different schools or classrooms, or are likely to undergo significant evolution during the field test."

2. Randomization Point

Where and when random assignment should occur depends on the particular school-to-work model or intervention being evaluated. For example, school-to-apprenticeship programs serving high school seniors presumably would randomly assign applicants prior to students' junior year based on their interest in enrolling in the program. However, under the career academy model, student applicants would have to be randomly assigned at the end of the eighth or ninth grade, depending on the length of the particular academy in question, based on students' desires to pursue such a path. Career magnet schools present a very different situation, one in which students would be randomized at their middle or junior high schools before the beginning of high school based on their choosing to enroll in a career-specific magnet high school. In this and similar cases, randomly assigning schools or matching them might be more appropriate, even though such approaches are generally not well received in the literature (e.g., Garfinkel et al., 1991). There are likely to be as many possibilities for random assignment points as there are school-to-work models.

3. Specific Outcome Measures

Researchers and practitioners have suggested numerous measures and indicators for gauging school-to-work performance, including both in-program and postprogram measures. Such measures go beyond the list specified in STWOA Section 402. Paul Barton (1994) developed a list of indicators of progress in school-to-work transition that included six measures of final outcomes, seven measures of intermediate outcomes, and ten systems outcomes. It should be pointed out that he focused on progress made by students who did not intend to enroll in college. Barton's recommended *final outcomes* for school-to-work included the following:

- Establishing employment earlier
- Reversing the relative earnings decline between high school graduates and school-to-work completers and two- and four-year college graduates
- Rewarding academic/literacy/skill achievement in the labor market
- Increasing economic independence
- Improving the perception of well-being among youth

His *intermediate outcomes* included the following:

- Increases in academic achievement and literacy
- Increases in the proportion of secondary school students in bona fide work-based learning programs using the worksite
- Increased enrollment in articulated school-to-work and Tech Prep programs

- Increases in awards of skill certifications
- Increases in the proportion of high school graduates who obtain their jobs with assistance from institutions involved in school-to-work
- Increases in employer training provided to 18- to 24-year-old employees
- Improvements in parental perception of the viability of the new school-to-work initiatives that are not identified as the traditional academiccollege preparatory route

An alternative list of potential school-to-work outcomes gleaned from the literature (e.g., Hoffinger and Goldberg, 1995; Bailis, 1995; Stern et al., 1995) includes the following items:

- Career awareness (e.g., higher proportion of students with a career goal who are knowledgeable about the preparation needed to achieve that goal)
- School attendance rates
- Rates of student disciplinary actions
- Proportions of students taking a more rigorous curriculum (especially in math, science, and technical courses)
- Various measures of student achievement
- Secondary school completion rates
- Entry rates to various forms of postsecondary learning
- Postsecondary credentials obtained
- Skill certifications obtained
- Higher earnings and improved employment stability on leaving school (e.g., earlier access to high-wage occupations, learning and advancement; reductions in "churning" or "job hopping"; steeper age-earning profiles indicating more rapid advancement with skill development; improvements in the quantity and quality of jobs attained by youth)

Ultimately, a smaller set of performance measures must be identified, defined, and measured if the evaluation's findings are going to be viewed as useful to policymakers and program administrators. These measures will vary with the interventions that are chosen as the basis for the evaluation of course. Not all measures are appropriate for evaluating all school-to-work models or treatments. This information again raises questions about what school-to-work is attempting to accomplish, whether broad-based school reform, improved transition to work, or enhanced access to quality employment, just to name a few of the many possible goals.

To the extent school-to-work emphasizes work or occupational outcomes, labor market measures will be much more important, whether they include employment rates, occupation of employment, starting and end-of-period wage rates, earnings, or similar measures. Both the quantity and quality of the employment outcomes for participating youth are important to measure for work-oriented programs. On the other hand, if improving employment outcomes is secondary to improving performance in school, then measures should reflect this emphasis. How precisely the selected work- and school-oriented outcomes need to be measured will have calculable effects on sample sizes and their allocation among treatment and control groups as well.

4. Target Populations or Subgroups

Along with increased variation in school-to-work models and treatments has come greater diversity in the groups being served. For example, while the early school-to-work initiatives attempted to serve almost exclusively in-school youth, more recent efforts have begun reaching out to serve out-of-school youth as well. School-to-work efforts aiming for broad-based school improvement must be viewed as encompassing more of the available student body. States and local school-to-work partnerships are now directed to serve *all* students, providing equal access to all program components for several subgroups mentioned specifically in the STWOA, including the following:

- Out-of-school youth (including dropouts)
- Low-income youth
- Low-achieving youth
- Limited English speakers
- Youth with disabilities
- Academically talented youth

- Youth in rural areas
- Racial/ethnic groups

The MDRC evaluation of career academies is examining impacts for three key student subgroups: (1) students at risk of educational failure as defined by the National Center for Educational Statistics (i.e., having three or more of the following six characteristics: student lives in a single parent household; student lives in a low-income household; student is home alone more than three hours a day; neither parent of student has a high school diploma; student has a sibling who dropped out of high school; student speaks limited English); (2) disengaged students (defined as having excessive absences, cuts, or disciplinary actions the semester prior to random assignment); and (3) overage students (students over the age for their grade level). While these categories are customized to accomplish the objectives of the MDRC study, they include three of the STWOA categories as components-low-income youth, low-achieving youth, and limited-English speakers.

Any school-to-work evaluation must be attentive to the relevant target population for the particular model or intervention being evaluated. Variation in school-to-work models and treatments thus may call for substantial target group variation as well. It may be that the only way to deal effectively with such wide variation is through several separate model-specific field studies, randomly assigning within given models or treatments and estimating impacts for key target groups. This study has important implications for site selection, sample size, and other related issues. To provide detailed, reliable impact estimates for all of the groups specified in the STWOA would be unrealistically expensive. The Department of Labor spent more than $20-plus million over seven years for the National JTPA Study, producing employment and earnings impact estimates only for adult men, adult women, and out-of-school youth in a limited number of treatment streams.

5. Site Selection

Site selection is an issue with several dimensions as well. Education systems in the United States are state-local systems funded predominantly by state and local sources. While some Federal aid flows to states and localities, the basic public education system is inherently state and local in nature.

Frameworks and institutional support for school-to-work within education systems vary a great deal among the states, territories, and Native American tribal nations.

Several school-to-work activities predated the passage of the STWOA in 1994. For example, the career academy movement traces its roots back to the late 1960s in Philadelphia. Even at the state level, initiatives were well underway. In 1992, the Council of Chief State School Officers, relying on grants from the Pew Charitable Trusts and others, worked to develop statewide youth apprenticeship systems with a network of eight states-California, Iowa, Maine, Michigan, Oregon, Pennsylvania, West Virginia, and Wisconsin (Council of Chief State School Officers, 1995). A number of states passed legislation and began their own school-to-work programs prior to enactment of the Federal legislation. These included Arkansas, Florida, Illinois, Indiana, Kentucky, Maine, Massachusetts, Michigan, New Jersey, New York, Oregon, Pennsylvania, and Wisconsin,

Since 1994, all states, the District of Columbia, Puerto Rico, and U.S. territories have received development grants to help them plan their school-to-worksystems. Grants for state implementation have been made in waves as states became ready for implementation. In addition, competitions were held for separate implementation grants for localities in states without implementation grants. Separate Federal competitions were held for partnerships serving high poverty areas and those serving Native Americans.

In the first round of competition for state implementation grants held in 1994, grants were awarded to eight states: Kentucky, Massachusetts, Maine, Michigan, New Jersey, New York, Oregon, and Wisconsin. Localities ready to implement school-to-work in states without implementation grants competed for direct Federal grants. A second state competition held in 1995 produced implementation grants for an additional nineteenstates-Alaska, Arizona, Colorado, Florida, Hawaii, Idaho, Indiana, Iowa, Maryland, Nebraska, New Hampshire, North Carolina, Ohio, Oklahoma, Pennsylvania, Utah, Vermont, Washington, and West Virginia. Thus, altogether as of February 1996, twenty-seven states were receiving Federal grants to implement state school-to-work systems. In addition, implementation grants have been made to seven territories: American Samoa, Northern Mariana Islands, Micronesia, Guam, Marshall Islands, Palau, and the Virgin Islands.

To qualify for award of a Federal implementation grant, states had to agree to promote and support statewide implementation of school-to-work programs, to plan to provide local funding to sustain the system after Federal funds sunset, and to coordinate activities from related Federal education and training programs, including the Carl D. Perkins Vocational and Applied Technology Education Act, the Elementary and Secondary Education Act, the Family Support Act, the Individuals with Disabilities Education Act and the Adult Education Act. In addition, states had to ensure access for all students and opportunities for young women to participate in programs leading to high-performance, high-paying jobs, including jobs in nontraditional areas.

At this writing, it is uncertain whether additional rounds of state implementation grants will be made, nor whether the states and territories with implementation grants will actually receive the full five years of Federal venture capital funding originally envisioned.

The nature and scope of the approach to school-to-work varies considerably by state, For example, Arkansas, Maine, Pennsylvania, and Wisconsin have focused on developing statewide youth apprenticeship systems. Oregon and Kentucky have endorsed forms of an occupational-academic cluster approach and in these states, school-to-work activities are built into a framework of broader school reform measures. North Carolina and Florida are noted for their strong Tech Prep systems.

There is also wide variation in implementation practices among *local* school-to-work programs. Programs/systems have multiple exit points (e.g., graduating from high school, obtaining a one-year community college certificate, receiving a two-year associate's degree in community college or four-year bachelor's degree, completing an apprenticeship). Another of the lessons cited by MDRC researchers was that "[n]arrow field tests that ignore the importance of local implementation issues don't work" (Pauly and Thompson, 1993). As support for this point, they cite the experience with the important Follow Through and Head Start Planned Variation studies that were reported on two decades earlier by Rivlin and Timpane (1975).

One way of dealing with substantial variation might be to confine the evaluation(s) to a few states or selected localities that are sufficiently comparable in terms of their school-to-work policies and practices. In any event, in view of the widespread state/local variation, data collection on participation in the various components will need to be performed meticulously.

Site selection for any net impact evaluation of occupational-academic clusters will be much more challenging than site selection was for the MDRC career academy study. In order to conduct a net impact study of occupational-academics clusters, care must be taken to select only suitable sites that are willing and fully able to participate. Only those schools and/or programs that are oversubscribed (or could become oversubscribed with more intensive recruiting) should be considered. This raises several possible complications. For example, one of the clusters may be oversubscribed while others accept all students who apply. Selection bias may be introduced in the control group if they are all rejected candidates from the popular cluster whereas the treatment group includes participants in several clusters. In addition, school districts with a single high school organized into occupational-academic clusters or schools in which all students participate and assignment to the school is based on residence should be avoided because they would offer no opportunities to obtain an appropriate control group using random assignment. Also, caution will be required where occupational-academic clusters are operated in multidistrict settings in which funds follow the learner. MDRC researchers discovered that financial arrangements often influence incentives for school districts to send students or to avoid sending them depending on whether its own enrollments are growing or declining (Pauly, Kopp, Haimson, 1994, p. 168).

Likewise, obtaining a sufficient population for study may pose problems in some of the youth apprenticeship programs, many of which are small. In order to conduct field trials using random assignment, the programs need to be oversubscribed so that a sufficient population is available for assignment to fill both the treatment group and the control group. In addition, random assignment is ethically only appropriate where the treatment is not an entitlement, but rather a form of treatment that needs to be rationed in some way because it is oversubscribed. School-to-work programs in some industries such as metalwork and construction reportedly have encountered difficulties in attracting sufficient numbers of applicants to fill their programs, let alone provide any extra candidates for a control group.

6. Sample Size and Representativeness

Determining the sample size for evaluations of this type depends on a number of design choices, including among others: the number of school-to-work models or treatment interventions for which impacts are being estimated; the desired precision of the impact estimates; minimum detectable impacts sought; expected sample no shows, attrition, and other rates; the number of target groups for which separate impacts are being estimated; and whether separate site effects are being estimated for the various school-to-work programs. These are relatively straightforward issues: once design decisions have been made on each of these other dimensions, computing the requisite sample sizes and their allocation becomes largely a known technical exercise.

There is also the issue of whether any resulting impact estimates should be nationally representative. Given the discussions concerning program variability at the state and local level, as well as the fact that school-to-work policies and programs are intended to become the domain of governors, and especially the emerging state and local partnerships, the evaluation design should probably forego attempts at national representativeness altogether. A bigger concern here may be whether it is possible to obtainvalid and reliable impact estimates for particular models or treatment interventions for key target populations, whatever their geographic area.

7. Timeframes

Evaluation timeframe issues include both those related to startup and to the duration of the program treatment or intervention (i.e., the expected time participants are expected to be receiving services), as well as to the period over which program results are likely to occur.

Evaluation Startup

Any evaluation involving random assignment should begin by including at least a one-year pilot phase in which site development (including mobilizing community support), technical assistance, and random assignment of participants and controls would occur.

Program Duration

The longer the duration of program interventions, the greater will be both "no-show" and attrition rates and the greater the risk of midstream changes in approach during the evaluation. According to the MDRC's preliminary findings, one-quarter of the students who originally enrolled in the career academies were not enrolled two years later. Mathematica reported slightly higher attrition in its study of youth apprenticeship programs across sixteen sites. The study, which included several new programs, compared initial enrollments in fall 1992 with numbers of students participating by the end of that school year in spring 1993. Among the thirteen programs for which data were available for comparable time periods, attrition averaged 17 percent, ranging from a high of 41 percent to zero (calculated from Mathematica Policy Research, 1994, p. 22). If such attrition rates persist throughout the program, from 13 to 17 percent of the participants would be lost each year.

A lengthy program duration also drives up the cost since the evaluation must stay in place to follow participants through the program and beyond the program to measure its effects on subsequent employment and earnings. Although some career awareness and career investigation activities are provided to students in elementary and middle schools, the average time in school-to-work is generally going to be around six years (from grades nine through fourteen).

Postprogram Time Period

While tracking student outcomes for one year following participation may be sufficient for some models or treatments, for others it may be far too short. For those models that aim to enhance the effectiveness of instruction within the schools or to ensure that all students are exposed to curriculum elements stressing knowledge of the labor market, very short followup tracking periods may be required. On the other hand, for some of the school-to-work models that are attempting to substantially alter the career possibilities and experiences of participating youth-for example, to gain them access to higher quality jobs, to reduce job changing, to increase their earnings over time-it may be necessary to track both the treatment and control group members for several years following program participation. Clearly these periods will range widely depending on the intervention in question. The longer the no-services embargo period, the less likely state or local programs may be interested in volunteering to participate in the evaluation (see Section IV).

Another effect of long program duration and postprogram measurement periods is likely to be increased crossover rates, that is, participants in one intervention then receiving services under another (unallowed) treatment, as well as controls ultimately securing services during the embargo period, thereby becoming treatment group members. Both results are damaging to the evaluation's impact estimates. On the one hand, treatment crossovers lend considerable ambiguity to the results and create some joint treatment effects that may not have been planned for, while treatment/control crossovers can be expected to impart downward bias to the impact estimates. The latter can be adjusted for to some extent.

The combination of long inprogram and postprogram timeframes may also mean that when the evaluation findings ultimately surface, they lose some of their currency and relevance for policymakers. With a one-year startup phase, six years of school-to-work treatment, and a followup period of three years beyond high school, a net impact evaluation begun in fall 1996 would not yield final outcomes data-let alone a report with findings and recommendations -until spring 2006 at the earliest. This is well beyond the September 30, 1998, due date for the evaluation contained in Section 402(b) of the STWOA. It is doubtful whether evaluation results retain their timeliness after a decade or more. Their policy shelf-life is likely to be far shorter.

IV. Implementation Issues

This section addresses selected implementation issues, drawing on the preceding design discussion and experience and lessons learned from the conduct of other net impact evaluations, including the evaluation of JTPA and MDRC's current evaluation of career academies. In many respects, it has been difficult to effectively separate evaluation design from implementation issues.

1. Treatment and Other Variation

That school-to-work models and treatments feature enormous within- and between-intervention variation, with many unknown hybrids of the programs operating in practice, makes conducting net impact evaluations of such efforts quite difficult. These problems are exacerbated by lack of consistent Federal component definitions and by the fact that state and local partnerships largely occupy the driver's seat for implementing these programs, resulting in considerable interarea variation as well.

Dealing with the problems introduced by such wide variation along so many dimensions is likely to be overwhelming. However, there are some options worth considering in implementing an evaluation or evaluations. First, it may make more sense from a policy and budgetary point of view to focus on evaluating a restricted set of school-to-work models or interventions and measuring their impacts very precisely.

Second, given the definite evolution and expansion that is underway within the school-to-work movement-embracing more school-reform models in addition to the initial emphasis on transitions from public education to the workplace-it may also be wise to invest scarce evaluation resources on producing net impact for models/interventions that are expected to remain relatively stable into the future. Otherwise, any findings are unlikely to be perceived as useful or relevant at any level. If so, policymakers will have spent a great deal of money for a report that will simply be occupying space on a shelf.

Third, school-to-work treatment and area variation dictates more strongly than in some earlier evaluations, that process and implementation studies be performed as essential complements to any net impact evaluations. These will be needed at the least to monitor deviations from planned program practices. In addition, for school-reform oriented school-to-work models, they may constitute the desired path-coupled with efforts to track gross postprogram outcomes-in and of themselves. Experimental evaluations featuring random assignment of students may not be appropriate for assessing the many possible and far-reaching impacts of school-to-work as public school reform.

Two other lessons that MDRC researchers gleaned from their evaluation efforts (Pauly and Thompson, 1994) should also be highlighted here, as follows:

- Careful choices of approaches to be field tested, high-quality data collection, sufficient numbers of participating schools and students, and in some cases long-term followup are necessary for field tests to be maximally useful.

- Different research designs are needed depending on whether the goal is to measure whether the impact of an intervention simultaneously affects an entire community, an entire school, or particular students within a school.

Note that not all variation is bad. That school-to-work models and programs vary widely in unplanned and unknown ways is what creates enormous difficulty for designing and implementing such an evaluation. However, to the extent that designs can take advantage of known or planned variation in existing programs, such evaluations may actually come out ahead. For example, if we know that one school-to-work model (or state) stresses much higher levels of employer participation-and it can be fully and consistently documented-then the evaluation may be able to statistically control for this participation and make inferences concerning the contribution made by such participation.

2. Lack of Standardized Definitions

The problems of variation are compounded by the lack of consensus on definition of terms. Although the National School-to-Work Office has tried to remedy this problem by promulgating a glossary of terms, the problem persists. For example, the current version of the glossary defines work-based learning as "Learning that takes place in the workplace." Yet, the intensity of a work-based learning experience may differ significantly. "Work-based learning" can mean a one-week job shadowing experience, a summer work-based learning experience, or the on-the-job portion of an apprenticeship extending over several years. Work-based learning also varies significantly in the extent to which it is tied into school work. Several other terms present equal or even greater difficulties. For example, the integration of academic and vocational education carries many different meanings.

3. Random Assignment

Given all of the other issues raised, random assignment of school-to-work students is likely to present serious barriers to implementing an impact evaluation. Not only is it going to be hard to determine where, when, and whom to randomly assign, but problems associated with instituting an experimental design over such a long span of time and in a school-based context may prove insurmountable for some of the school-to-work program models, MDRC's early success with the career academies evaluation notwithstanding.

Generally, random assignment should be conducted as near to the point of enrollment as possible. Who qualifies as a participant in school-to-work programs and at what point do they qualify? The legislation states that students should have career information "no later than seventh grade" and that they should select a career major "no later than eleventh grade." If an evaluation begins with seventh graders and follows them through one or two years of postsecondary education or training, this means seven years in the program plus a followup period after leaving school.

For school-to-work systems, this is likely to be eighth to tenth grade (or thirteen to sixteen year-olds). This means that obtaining informed parental consent will be needed to gain access to school records (including attendance, achievement, course work, graduation status, and disciplinary actions). With as much fanfare as some of the school-to-work models have been introduced in local communities, what well-informed parent is going to voluntarily sign away his child's rights to participate in such an initiative for the present and for the next five, six, or more years? Desirable alternative approaches are not readily apparent, unless the quasi-experimental matching procedures described by Heckman prove feasible for youth as well as adults.

4. Participation Incentives

Participation incentives may pose special problems, especially since Federal funding for school-to-work may cease altogether in FY '97. What incentives should be made available to overcome the extra burdens of putting up with program data collection and research and dealing with the extended embargo issues? Less severe versions of these issues presented themselves in the national JTPA study (Doolittle and Traeger, 1990). That evaluation had to cope with a few added constraints in the beginning, including the goal (later dropped) of securing a nationally representative sample of programs, grouped according to size and urban/rural setting, as well as circumventing difficulties involved with performance incentive grants potentially foregone by local JTPA programs if their governors were unwilling to exempt them from performance standards. The existing school-to-work program environment is not encumbered by the latter problem. An evaluation of school-to-work programs need not worry so much about securing nationally representative sites; rather, the primary focus should be on measuring very well what is being measured.

The MDRC evaluation of career academies was able to attract school districts to participate by giving them grants to offset the burden of added data collection for the project and by promising them networking opportunities in conferences of all the study sites held annually during the study. In addition, a major motivation for educators was the prospect of obtaining some definitive data on whether the career academy model was effective.

To the extent that treatment and followups are of lengthy duration, the no-services embargo periods will have to be extreme in length. States and localities are not likely to volunteer to participate—nor are students and their parents likely to give their informed consent to participate as possible control group members-without the provision of substantial incentives. It is not at all clear what could be provided to the families in this regard without seriously damaging the evaluation. There is also the concern that overly generous financial incentives for participation could have the effect of distorting local school-to-work system or program performance during the course of the evaluation.

5. Intervention and Outcomes and Duration

Given the length of students' expected participation in school-to-work interventions-some for as long as four to six years-sample attrition and crossover issues are likely to be even greater than in any prior program evaluations. Most evaluations such as the national JTPA study (e.g., Doolittle and Traeger, 1990; Bloom et al., 1993) tracked program participation for adults and out-of-school youth for periods of about fifteen to eighteen months with additional postprogram tracking time. Yet, school-to-work interventions are likely to be far longer before postprogram outcome tracking can begin, followed by embargo enforcement for many years.

Obtaining adequate followup (outcomes) data on students beyond high school can be both difficult and expensive. One way to minimize the difficulty and reduce the cost, at least in terms of employment-related outcomes, is to make extensive use of unemployment insurance (UI) wage records, which can be obtained in all the states with state school-to-work implementation grants except New York, which still relies on state income tax records. However, as several National Commission for Employment Policy reports (NCEP 1991, 1992) have clearly documented, UI wage records have important limitations for purposes of evaluating school-to-work employment and earnings outcomes as well, including the following:

- A state's UI wage records do not provide information on out-of-state employment. This is less of a problem than it used to be in that, with encouragement and support from the national UI office, the various state UI offices have become skilled at working out interstate data sharing agreements.

- UI coverage also has a few holes, though these have become quite small over the years: in Texas, for example, UI coverage of wage and salary employment is estimated to exceed 98 percent.

- UI records also lack variables for employment start and end dates and hourly wage rates among other key shortcomings.

- Occupational or job-title codes are only available on UI wage records in a limited number of states; for example, Alaska for all UI-covered employment and Florida and Texas for employment encompassed by their student and job training participant followup systems (King and Lawson, 1996). The lack of occupational information seriously limits the analysis on the quality of jobs secured by school-to-work participants.

- Some youth may be employed by temporary employment agencies, which are classified in the Standard Industrial Classification (SIC) code for business services rather than the actual industry of employment. This shortcoming has far broader effects than just school-to-work evaluation implementation.

- Analysis of job hopping may be hampered by the issue of mergers and acquisitions that result in changes of the employer's SIC code without the worker actually switching employers. This factor tends to overstate job hopping and make it difficult to obtain reliable and consistent employment change estimates for youth exiting school-to-work work-oriented models.

In general, while UI wage records can supply much if not most of the postprogram labor market outcomes information desired, they have their shortcomings and will probably need to be supplemented through surveys with former students and their employers. These surveys clearly add to the cost of the evaluation.

6. State and Local Partnerships: The Key Audience

One very important feature of school-to-work program implementation is its heavy reliance on state and local partnerships. As indicated above, in a very short time, the Federal roles of offering policy guidance and technical assistance and of providing the seed money for these programs are likely to diminish if not disappear altogether. If evaluations of school-to-work programs are going to be performed well and their findings perceived as useful to policymakers, program planners, and administrators down the road, these very state and local partnerships should play a central role in virtually every aspect of their design and implementation. This means contributing in all likelihood the majority of their funding.

V. Conclusions

The STWOA calls for the Secretaries of Education and Labor, in collaboration with the states and localities, to complete a national evaluation of programs funded under the Act to assess the implementation of programs and their effectiveness and to establish a system of performance measures for assessing state and local programs (Section 402). The legislative language is relatively open regarding what evaluation methodologies are to be used. The Departments of Labor and Education and the National School-to-Work Office are currently examining the feasibility of using a net impact evaluation of the programs on student participants using a randomized experimental design. The evaluation literature is generally supportive of such an approach, but there are limitations and conditions under which such a study may be inappropriate. For example, a net impact evaluation of school-to-work activities on students is probably not the appropriate tool to assess systems changes in reforming public schools. Also, while focusing the net impact of school-to-work programs on students, we hasten to mention that another important dimension not addressed is the impact on employers (in improving the quality of job applicants or entrants available to them, improving the competitiveness of participating firms, and providing long-run return-on-investment).

Clearly, having solid evidence regarding the effectiveness of school-to-work initiatives on students is desirable. The request for evidence made by the superintendent described in the introduction to this paper needs to be answered, especially for the many state and local partners who are putting so much energy and resources into this endeavor, Ultimately, such evidence will be even more critical to the future of the school-to-work movement in the coming block grant environment.

While we agree that a net impact study of school-to-work is desirable, this paper raises serious questions about its feasibility and appropriateness under the present circumstances. MDRC's Pauly and Thompson (1993, p. 23) concluded as follows:

> Information from random assignment field tests of interventions . . . would not provide policymakers with clear information on particular, clearly specified organizational structures or educational methods, and this would presumably limit the studies' value. A better approach might be to study the implementation of these interventions, documenting the methods used and the reasons that these methods were used. These implementation studies might result in hypotheses about particular substantive interventions for which a random assignment field test would be appropriate.

At this time, any proposed net impact evaluation of school-to-work would need to address at least two fundamental challenges: program variability, and program instability.

1. Program Variability

There is enormous variety in school-to-work programs across the nation, stemming from several sources, including the evolving vision of the school-to-work movement, differences in state frameworks and approaches, program treatments, models, implementation practices, and targeted subgroups. Many school-to-work activities predated passage of Federal legislation, and the STWOA program components have been implemented, not as prescriptive mandates, but rather as guiding principles. Customization to fit the needs and circumstances of localities, industries, and youth populations has been wisely encouraged. However, such variety creates problems for conducting a good net impact evaluation.

While some planned variation and site-specific differences can enrich an evaluation, unplanned and unknown (and unknowable) variations can destroy it. Also, there are too many models and treatment components and variations of models and treatments, too many outcomes, and too many target groups to be accommodated in an evaluation with an affordable price tag. Any evaluation undertaken should attempt to measure limited number of program impacts for only those target groups and interventions/models that are most likely to remain the ones of greatest policy concern down the road in states and localities-the primary audience for this evaluation. A further implication is that, rather than conducting any sort of overall school-to-work evaluation, the design should be geared to conduct model-by-model or treatment stream field studies.

2. Program Instability

Any program cannot be evaluated properly if it is not at least semimature and somewhat stable. Given the changes in funding levels, shifts in administration at the Federal, state, and local levels, and the continual evolution of program vision and content, it seems unlikely that the school-to-work initiative will remain intact and stable over the lengthy period required for a net impact evaluation. The treatment period for school-to-work programs is more lengthy than experienced in any previous net impact study, and the overall period of evaluation called for would likely be longer than any previous study, with the possible exception of the Perry preschool evaluation.

3. Opportunities for Action

The Federal government needs to build on the work already under way. Efforts to develop a glossary and conventional definitions of terms and a consensus set of progress measures through the National School-to-Work Office are worthwhile endeavors and need to be continued and refined. MDRC has successfully mounted a study of career academies, using random assignment techniques (Kemple and Rock, 1996). The career academy is perhaps the tightest and most promising of the school reform school-to-work models in place today. Additional funding is needed to complete the followup portions of this study should the initial findings warrant. Likewise, the current evaluations undertaken by Mathematica need to be followed through to completion, providing evidence on gross outcomes and implementation practices that could lead to a useful net impact study at some future point.

Any net impact evaluation undertaken needs to be conducted in a selected set of sites that have a very tight definition of interventions and that are willing to hold to them over long timeframes. Probably the only way to accomplish this impact is though a tightly controlled demonstration project undertaken and funded collaboratively with states. As part of the evaluation, the project should prepare the site staff well and implement training across sites to help ensure integrity of treatment design through the course of the evaluation. Any net impact evaluation should be accompanied by a careful process evaluation in all the sites.

Evaluation of School-to-Work Transition Programs

James J. Heckman
Henry Schultz Distinguished Service Professor
University of Chicago

I. Introduction and Summary

This paper considers alternative strategies for evaluating school-to-work transition programs. Randomized social experiments are compared to nonexperimental evaluation methods such as instrumental variable methods, differences-in-differences methods, traditional selection methods, and matching methods. This paper presents empirical lessons from the concluding evaluation of the Job Training Partnership Act (JTPA) sponsored by the Department of Labor, the Russell Sage Foundation, and the National Science Foundation.

This paper starts from the following point of view. An ideal social experiment that does not disrupt the program being evaluated does many things at once: (1) it places treatments and controls in the same local labor market, (2) it administers **the same** questionnaire to both groups, (3) it balances selective differences in unobserved characteristics between control and treatment groups, and (4) it balances the distribution of characteristics between the groups.

Using data from the JTPA evaluation and a collection of nonexperimental comparison groups and, in particular, focusing on the evaluations for young people, we ask which of these features is the key to the success of the experimental method. The surprising answer is that Factor **3--selection bias--** defined as selective differences in unobservables-plays only a very modest role in accounting for differences between experimental controls and members of comparison groups of nonparticipants (persons who did not apply to the program). It plays a substantially larger role for youth when **no-**shows (persons who were accepted into the program but who did not show up to participate) are used as a comparison group. In the instance of the JTPA program, the success of the experimental method arises more from matching people on the same characteristics, placing them in the same labor market, and administering them the same questionnaire. Similar results are found in our analysis of the National Supported Work data previously analyzed by Fraker and Maynard (1987) and LaLonde (1986).

This evidence suggests the ingredients required for a successful nonexperimental evaluation of job training programs: match people in the same labor market, use the same questionnaire and definitions of variables, and balance the distribution of observed characteristics. The only other required piece of information is knowledge of the determinants of who participates in the program (for a study of a comparison group based on nonparticipants) or knowledge of the determinants of who fails to show up (for a study of no-shows). With these ingredients, nonexperimental methods produce estimates of program impact that are extremely close to experimental estimates for a variety of demographic groups and social experiments.

The structure of the paper is as follows. Section II presents the instrumental variable estimator for a general model of program evaluation. Implicit behavioral assumptions underlying the method are discussed. Section III discusses the experimental method and shows how randomization acts as an instrumental variable. An important contrast is drawn between the conventional interpretation of instrumental variables made in the classical econometrics literature and the use of instrumental variables in evaluation research.

Several points of randomization are discussed: (1) randomization at the point in the enrollment process where persons apply and have been accepted into the program and (2) randomization of eligibility for the program. Neither type of randomization estimates the effect of the program on persons selected at random from the population to participate in it. Both are designed to estimate the effect of "treatment on the treated."

Section IV discusses the method of matching. The sources of differences between treatment and comparison group members are decomposed into the four distinct components previously discussed. Selection bias is a fairly minor part of the problem in both the National Supported Work and JTPA data. Lessons for the nonexperimental evaluation of school-to-work transition programs are extracted. The paper concludes with a summary of the main arguments.

II. The Method of Instrumental Variables

1. Introduction

This section of the paper considers the use of instrumental variables to estimate the mean effect of treatment on the treated. I establish that in cases in which the responses to the treatment vary, the instrumental variable argument fails unless person-specific responses to treatment do not influence the decision to participate in the program. This requires that individual gains from the program that cannot be predicted from variables available to observing social scientists do not influence the decision of the persons being studied to participate in the program. In the likely case in which individuals possess private information about gains from the program that cannot be proxied by the available data, instrumental variables methods do not estimate behaviorally interesting evaluation parameters. Instrumental variable models are extremely sensitive to assumptions about how people process information.

The method of instrumental variables is now widely used in evaluation research. It has recently been promoted by **Angrist**, Imbens, and **Rubin** (1996), but was first discussed in the context of evaluation research by **Heckman** and Robb (1985, 1986). To understand the method, consider the following commonly used model for evaluating programs. Let Y be an outcome such as earnings. Let $D = 1$ if a person participates in a training program, $D = 0$ otherwise. Using a standard regression model, write

$$Y = \alpha + \beta D + U$$

where U is an error term with mean zero ($E(U) = 0$). This term arises from omitted variables that may determine the outcome. Much concern has been expressed about the possibility that D is correlated with U. Persons who would have high values of U (or Y) in the absence of the program may be the ones who go into it. This could happen if there is "cream-skimming" in which program officials select persons with high U. Formally, the fear translates into the concern that

$$E(U \mid D=1) \neq 0.$$

In this example, this term would be positive. In this case, simple regression estimates of β are upward biased.

An instrumental variable Z has two properties. It is uncorrelated with U:

(1) $E(U \mid Z) = 0.$

It is correlated with D:

(2) $E(D \mid Z) = \Pr(D=1 \mid Z)$ is a nontrivial function of Z taking distinct values for at least two distinct values of Z.

Applying these properties to the outcome equation, take expectations conditional on Z:

$$E(Y \mid Z) = \alpha + \beta \Pr(D=1 \mid Z).$$

Using the assumption that there are two distinct values of Z, say Z_1 and Z_2, with distinct probability values (i.e., distinct values of $\Pr(D=1 \mid Z)$) one can solve for β

$$\begin{aligned} E(Y \mid Z_1) &= a + \beta \Pr(D = 1 \mid Z_1) \\ E(Y \mid Z_2) &= \alpha + \beta \Pr(D = 1 \mid Z_2) \end{aligned}$$

$$\beta = \frac{E(Y \mid Z_1) - E(Y \mid Z_2)}{\Pr(D = 1 \mid Z_1) - \Pr(D = 1 \mid Z_2)}.$$

Using sample moments to replace population moments, one obtains the instrumental variables estimator.

The method clearly requires an instrument Z. How credible are the instruments? Under what conditions does the method work? Can the method be generalized to cover the plausible case where β differs among people, i.e., there is variable response to treatment instead of a common response for everyone?

This section of the paper addresses these questions. I establish that in cases in which the responses to the treatment vary, the instrumental variable argument fails unless person-specific responses to treatment do not influence the decision to participate in the program. This requires that individual gains from the program that cannot be predicted from variables available to observing social scientists do not influence the decision of the persons being studied to participate in the program. In the likely case in which individuals possess private information about gains from the program that cannot be proxied by the available data, instrumental variables methods do not estimate interesting evaluation parameters. Instrumental variable models are extremely sensitive to assumptions about how people process information.

The intuition behind this result is that when responses to a program differ among persons, the effect of the program on raising participant outcome measures compared to what they would be in the absence of the program depends on the composition of persons in the program unless the effect of the program is the same for everybody or else persons are somehow randomly sorted into the program with respect to outcomes. One way people can be randomly sorted into the program is if they do not know their own gain from the program at the time they enter it and they cannot forecast the gain from the program or any of the components at that time.

If people can forecast the gain, or components of the gain, then the instrumental variable estimator does not estimate the impact of participation for participants. This is so because in the case of heterogeneous impacts, β varies among people and the evaluation parameter comparable to β in the simple case previously discussed depends on Z for the population of persons who go into the program $(E(\beta \mid Z, D=1)=\beta(Z))$. Then the instrumental variable estimation strategy breaks down because $E(Y \mid Z) = \alpha+\beta(Z)\Pr(D=1 \mid Z)$.

Thus,

$$E(Y \mid Z_1) = \alpha + \beta(Z_1)Pr(D=1 \mid Z_1)$$
$$E(Y \mid Z_2) = \alpha + \beta(Z_2)Pr(D=1 \mid Z_2)$$

and it is no longer true that we can use simple algebra to derive $\beta(Z_1)$ or $\beta(Z_2)$ i.e.,

$$\beta(Z_1) \neq \frac{E(Y \mid Z_1) - E(Y \mid Z_2)}{\Pr(D=1 \mid Z_1) - \Pr(D=1 \mid Z_2)} \neq \beta(Z_2).$$

However, $E(Y \mid Z)$ can always be estimated. Variation in Z informs us about how the "total gain" minus the product of the mean gain (p(Z)) and the participation rate ($\Pr(D=1 \mid Z)$) vary, although it does not identify the mean gain for those who go into the program.[1]

The intuition for why β depends on Z in the case of anticipated heterogeneous responses to the program is as follows. "Z" consists of variables that determine who goes into the program. We observe Z and so does the program participant. But he also observes a person-specific gain, or a component of that gain. The observing economist does not. In the draft lottery, numbers were listed in decreasing order of the likelihood that the person will be drafted. The lottery number is the instrument. Suppose for example Z is a variable that decreases the probability of program participation. A person with a very high value of Z is unlikely to be drafted. For that person still to serve in the military, the gain from doing so must be high. Thus, among those who choose to enter the program, $\beta(Z)$ is higher the higher the value of Z.

If β is the same for everyone or if the realized gain from the program cannot be forecast at the time program participation decisions are made, there can be no tradeoff between gains and costs. Instrumental variables methods are valid unless Z determines β for other reasons. For a potential draftee, a person with a high Z may be a better training risk because he is less likely to be inducted into the army. Employers may invest more in such persons and for that reason their wages may be higher. For these cases, Z is not a valid instrument, but the reason is different. Now Z directly determines β even before conditioning on participation in the program.

---

[1] More formally, the effect of Z on Y can be estimated. Thus, we can identify the derivative $\left[\frac{\partial E(Y \mid Z)}{\partial Z} = \beta(Z)\frac{\partial \Pr(D=1 \mid Z)}{\partial Z} + \frac{\partial \beta(Z)}{\partial Z}\Pr(D=1 \mid Z)\right]$ We can also estimate $\frac{\partial E(Y \mid Z)}{\partial Z} / \frac{\partial \Pr(D=1 \mid Z)}{\partial Z} = \beta(Z) + \left(\frac{\partial \beta(Z)}{\partial Z}\right) / \left(\frac{\frac{\partial \Pr(D=1 \mid Z)}{\partial Z}}{\Pr(D=1 \mid Z)}\right).$

In order to establish these points in a more formal way, it is necessary to step back from the standard regression model and consider a more general model that captures the notion of heterogeneous responses.

2. The Evaluation Problem

The evaluation problem is a missing data problem. Persons may be in either one of two states but not both at the same time. The states are denoted "0" and "1," respectively. Outcomes are (Y_0, Y_1). Let $D = 1$ if a person is in state "1"; $D = 0$ otherwise. The outcome observed for an individual, Y, is

$$Y = DY_1 + (1-D)Y_0$$

This is an instance of the Roy model (1951) or a switching model (see Goldfeld and Quandt, 1972). Statisticians call this the "Rubin model" after one clear exposition of Fisher's model of experiments set forth by Rubin (1978). If "0" is the no program state, and "1" is the program state, the gain to participating in the program is

$$\Delta = Y_1 - Y_0$$

If, contrary to hypothesis, we could simultaneously observe Y_1 and Y_0 for the same person, there would be no evaluation problem. One could construct Δ for everyone.

To cast the model into familiar econometric notation, write the two-population model of Fisher (1951), Cox (1958), Roy (1951), or Rubin (1978) as a function of observables (X) and unobservables (U_1, U_0):

$$Y_1 = g_1(X) + U_1 \tag{1(a)}$$

$$Y_0 = g_0(X) + U_0 \tag{1(b)}$$

where

$$E(U_1)=0=E(U_0).$$

It is assumed that g_1 and g_0 are nonstochastic functions. For the familiar case of linear regression, the g functions specialize to

$$g_1(X) = X\beta_1$$
$$g_0(X) = X\beta_0$$

There are many forms of the evaluation problem depending on what feature of the missing data one seeks to construct. The most common form of the problem is cast in terms of means. One mean that receives the most attention is given below:

The Mean Effect of Treatment on the Treated

$$(2) \qquad E(Y_1 - Y_0 \mid X,D=1) = E(\Delta \mid X,D=1)$$

$$= g_1(X) - g_0(X) + E(U_1 - U_0 \mid X,D=1).$$

This mean answers the question, "How much did persons participating in the program benefit compared to what they would have experienced without participating in the program?" It is a nonstandard parameter from the vantage point of conventional econometrics because it combines "structure" (the g_0 and g_1 functions) with the means of error terms (U_0 and U_1).[2]

A second mean also receives some attention in the literature-the effect of randomly selecting persons from the general population into the program, which is given below:

Mean Effect of Treatment Randomly Applied to the Population

$$E(Y_1 - Y_0 \mid X) = E(\Delta \mid X) = g_1(X) - g_0(X) + E(U_1 - U_0 \mid X).$$

---

[2] Heckman and Robb (1985, 1986) present conditions for identifying this parameter using instrumental variables in nonexperimental settings. Their conditions apply to the general "variable treatment effect case" of equations 1(a) and 1(b). See also Heckman (1995) for the implicit behavioral assumptions invoked in using instrumental variables to estimate parameter (2) when responses to treatment are heterogeneous.

This mean answers the question of how much the average income would be affected if participation in a program was universal, assuming that there are no general equilibrium effects. Alternatively, this parameter is the effect of taking a person from the general population at random and moving him or her from "0" to "1". Further discussion of these parameters and their relationship to the traditional parameters of cost-benefit analysis is presented in **Heckman** and Smith (1995). Although the assumption of separability between X and U is conventional in econometrics, it is not required to define $E(Y_1 - Y_0 \mid D = 2,X)$ or $E(Y_1 - Y_0 \mid X)$, nor is it necessary to assume such separability in deriving estimates from experiments.

For certain purposes, it is also of interest to inquire about distribution of gains: $F(A \mid D=1,X)$ or $F(A \mid X)$. But it has been shown that social experiments, unaccompanied by further assumptions, cannot recover these distributions (see **Heckman,** 1992, or Clements, **Heckman,** and Smith, 1997). Under ideal conditions, social experiments recover $F(Y_0 \mid D=1,X)$ and $F(Y_1 \mid D=1,X)$ if randomization is administered at a stage of the application and acceptance decision at which persons would ordinarily be accepted into programs, and there is no attrition from the program.

The evaluation problem arises from the fact that ordinary observational data do not provide sample counterparts for the missing counterfactuals. For means, experiments supplement observational data by providing the information needed to form the sample counterpart of $E(Y_0 \mid D=1,X)$. More generally, social experiments supplement observational data and produce the information needed to form the empirical distribution counterpart of $F(Y_0 \mid D=1,X)$. Randomization provides the sample counterparts to these population objects if randomization bias induced by the process of experimentation is assumed to be unimportant (Heckman, 1992).

This intuitively appealing counterfactual is very difficult to estimate. Picking a millionaire at random to participate in a training program for low skilled workers may be an intriguing thought experiment but is neither policy relevant nor feasible. It is not policy relevant because interest centers on the effects of programs on intended recipients-not on persons for whom the program was never intended.

2. Constructing Counterfactuals

How does one go about constructing counterfactuals? That is the topic of a vast literature. Much of the literature on constructing counterfactuals draws on econometric models.

Consider a model in which the outcomes depend on explanatory variables X. In the traditional regression setting,

3(a) $$Y_0 = X\beta_0 + U_0$$

and

3(b) $$Y_1 = X\beta_1 + U_1$$

where $E(U_0 \mid X) = 0$ and $E(U_1 \mid X) = 0$. Nothing said here relies on linear regression models. We can just as easily use more general models, and we do so in Appendix A, which develops a more general, nonparametric framework.

The observed outcome Y can be written as

$$Y = DY_1 + (1-D)Y_0$$

which when 3(a) and 3(b) are substituted into it becomes

(4) $$Y = X\beta_0 + D[X(\beta_1 - \beta_0) + U_1 - U_0] + U_0.$$

This is a "two regime" or "switching" model, sometimes known as the Roy model (Heckman and Honoré, 1990).

The term in brackets multiplying D is the gain from moving from "0" to "1." The average gain for a randomly selected person in the population is $X(\beta_1 - \beta_0)$. The idiosyncratic gain is $U_1 - U_0$. In this notation the average gain is

$$E(\Delta \mid X) = X(\beta_1 - \beta_0).$$

The effect of treatment on the treated is

$$E(A \mid X, D=1) = X(\beta_1 - \beta_0) + E(U_1 - U_0 \mid X, D=1).$$

This expression differs from the former by the additional term $E(U_1 - U_0 \mid X, D=1)$. This tells you how much the gain of participants differs from the average gain that would be experienced by the entire population. This is the gain to the movers from going from "0" to "1."

We may rewrite equation (4) in terms of these two parameters. Thus, we may write the average gain parameter as a parameter of the following equation:

$$(5) \qquad Y = X\beta_0 + D[E(\Delta \mid X)] + \{U_0 + D(U_1 - U_0)\}.$$

We may rewrite (4) in terms of the parameter "treatment on the treated" by

$$(6) \qquad Y = X\beta_0 + DE(A \mid X, D=1) + \{U_0 + D[U_1 - U_0 - E(U_1 - U_0 \mid X, D=1)]\}.$$

Simplifying back to a familiar notation we can think of $X\beta_0$ as the "intercept" $\alpha(X)$ of the regression where the intercept depends on X. The gain is $\beta(X) = X(\beta_1 - \beta_0) + U_1 - U_0$. We write α and β only as functions of observables, X. The mean gain conditional X is

$$\bar{\beta}(X) = E(\beta(X) \mid X) = E(\Delta \mid X) = X(\beta_1 - \beta_0)$$

The idiosyncratic component of the gain is $\epsilon = U_1 - U_0$. Then we can write (4) as

$$Y = \alpha(X) + D\beta(X) + U.$$

It is notationally more convenient to keep the dependence of α and β on X implicit so the preceding expression becomes:

$$Y = \alpha + D\beta + U.$$

One may write equation (5) as

(5′) $$Y = \alpha + D\bar{\beta} + \{U + D\epsilon\}.$$

The effect of treatment on the treated is β^* where

$$\beta^* = E(A \mid X,D=1) = E(\beta \mid X,D=1) = \bar{\beta} + E(\epsilon \mid X,D=1).$$

In this notation, equation (6) may be written as

(6′) $$Y = \alpha + D\beta^* + \{U + D(\epsilon - E(\epsilon \mid D=1))\}$$

where conditioning on X is kept implicit.

Can we run a regression of Y on D with an intercept to estimate β^* and $\bar{\beta}$? The regression coefficient for D can always be written as the difference between two means. Let β be the probability limit-the value of the regression coefficient in large samples under conventional assumptions. There are three alternative representations of this limit.

(7a) $$\tilde{\beta} = E(Y \mid X,D=1) - E(Y \mid X,D=0)$$

(7b) $$\tilde{\beta} = \bar{\beta} + E(\epsilon \mid X,D=1) + [E(U \mid X,D=1) - E(U \mid X,D=0)]$$

(7c) $$\tilde{\beta} = \beta^* + E(U \mid X,D=1) - E(U \mid X,D=0)$$

From (7b), we see that β is biased for $\bar{\beta}$ by an amount

$$E(\epsilon \mid X,D=1) + [E(U \mid X,D=1) - E(U \mid X,D=0)]$$

From (7c), we see that $\tilde{\beta}$ is biased for β^* by an amount

$$E(U \mid X,D=1) - E(U \mid X,D=0) \tag{8}$$

Term (8) is sometimes called the ***selection*** bias ***term***. It tells us how the outcome in the base state differs between program participants and nonparticipants. Such differences cannot be attributed to the program. In terms of the previous notation and recalling that $U_0 = U$,

$$\begin{aligned} E(U \mid X,D=1) - E(U \mid X,D=0) &= E(U_0 \mid X,D=1) - E(U_0 \mid X,D=0) \\ &= E(Y_0 \mid X,D=1) - E(Y_0 \mid X,D=0) \end{aligned}$$

The difference between $\bar{\beta}$ and β^* is the difference between the unobservable gain for an average person in the population (defined to be zero) and the unobservable gain for the average participant:

$$E(\epsilon \mid X,D=1) = E(U_1 - U_0 \mid X,D=1).$$

β^* differs from $\bar{\beta}$ by the amount of the gain in unobservables between the two distributions for those who make the move. These unobservables may be observed by the person or persons deciding to go into the program. They are unobserved only from the point of view of the social scientist trying to estimate the impact of the program. They coincide when the mean gain in the unobservable conditional on D is zero i.e.,

$$E(\epsilon \mid X,D=1) = 0 = E(U_1 - U_0 \mid X,D=1).$$

We now present two special cases when $\bar{\beta}=\beta^*$.

3. When Does $\bar{\beta}=\beta^*$?

There are two important special cases when $\bar{\beta}=\beta^*$. The first is when $U_1 = U_0$ In this case, there are no unobservable components of the gain. This model—called the "dummy endogenous variable model" (see, e.g., Heckman, 1978)--is widely used in applied work (see Ashenfelter, 1978, LaLonde, 1986). It assumes that conditional on X, the effect of program participation is the same for everyone. This is sometimes called the common coefficient model.

The second case is more subtle: $U_1 \neq U_0$. But $U_1 - U_0$ or information correlated with or dependent on it does not determine who goes into the program. Suppose, for example, that at the time people go into the program they do not know $\epsilon = U_1 - U_0$ Their best forecast of ϵ is zero or some other constant. Then if their experience of ϵ is typical of that of the entire population, $E(\epsilon \mid X,D=1) = 0$, and $\bar{\beta} = \beta^*$. This case shares many features in common with the "random coefficients" model of traditional econometrics. This comparison is pursued in Appendix B.

Observe that in either the case where $U_1 - U_0 = 0$ or the case where ϵ is not forecastable at the date enrollment decisions are made, the problem of estimating $\bar{\beta}$ or β^* using the difference in outcomes between participants and nonparticipants, comes down to the problem arising from D being correlated with, or stochastically dependent on, U.

Also note that for the estimation of (6'), from the definition of β^*, D is constructed to be uncorrelated with $D(\epsilon - E(\epsilon \mid X,D=l))$. Thus, *irrespective of whether* $E(\epsilon \mid X,D=l)=O$ or not

$$E(\epsilon - E(\epsilon \mid X,D=1) \mid X,D=1) = E(\epsilon \mid X,D=1) - E(\epsilon \mid X,D=1)=0$$

Thus, even in the case where participation in the program is made at least in part on unobservable components of gain, the only source of correlation between the "error term" and the treatment variable *"D"* arises from U, if the coefficient of interest is β^*, the effect of treatment on the treated.

4. The Method of Instrumental Variables

A standard method for estimating parameters in econometrics is the method of instrumental variables discussed in the introduction. Here, we consider application of the method in a more general context. Instrumental variables must satisfy two basic conditions and a third derived condition. They are mean-independent of the error terms of equations (5') and (6'), i.e.,

(C-1-a) $$E(U + D\epsilon \mid X,Z) = 0$$

for identifying $\bar{\beta}$ or

(C-1-b) $$E[U+D(\epsilon - E(\epsilon \mid D=1)) \mid X,Z] = 0$$

for identifying β^* A second condition is that D depends on Z in the following way:

(C-2) $$E(D \mid X,Z) = \Pr(D=1 \mid X,Z)$$

is a function of Z. Z is not fully explained by X. Implicit is the assumption that the probability is defined for two or more values of Z, so that the probability of participation is a nontrivial function of Z, and the values of the probability differ for different Zs at least for some X.

The important idea in this condition is that the probability of participation depends on Z as well as on X. A third condition is that the dependence of Y on Z operates only through D. This condition is really a consequence of the first two conditions, but its role is so central that I break it out as a separate condition. Reminding the reader that α, $\bar{\beta}$ and β^* may depend on X, for $\bar{\beta}$ the derived third assumption is

(C-3-a) $$E(Y \mid X,Z) = \alpha + \bar{\beta}E(D \mid X,Z) = \alpha + \bar{\beta}\Pr(D=1 \mid X,Z).$$

For β^* the assumption is

(C-3-b) $$E(Y \mid X,Z) = \alpha + \beta^*E(D \mid X,Z) = \alpha + \beta^*\Pr(D=1 \mid X,Z).$$

For two distinct values of Z, say Z_1 and Z_2, such that $\Pr(D=1 \mid X,Z_1) \neq \Pr(D=1 \mid X,Z_2)$, we can identify $\bar{\beta}$ or β^* by forming a simple mean difference. Thus, from (C-3-a), (C-2), and (C-l-a),

$$E(Y \mid X,Z_1) - E(Y \mid X,Z_2) = \bar{\beta}(\Pr(D=1 \mid X,Z_1) - \Pr(D=1 \mid X,Z_2))$$

so

$$\bar{\beta} = \frac{E(Y \mid X, Z_1) - E(Y \mid X, Z_2)}{\Pr(D=1 \mid X, Z_1) - \Pr(D=1 \mid X, Z_2)}.$$

Replacing population means with sample means produces the instrumental variables estimator which, under standard conditions, converges to p(X). By similar reasoning, using (C-3-b), (C-2), and (C-l-b) to obtain,

$$\beta^* = \frac{E(Y \mid X, Z_1) - E(Y \mid X, Z_2)}{\Pr(D = 1 \mid X, Z_1) - \Pr(D = 1 \mid X, Z_2)}, \tag{9}$$

Loosely speaking, instruments are variables that "don't belong in the population outcome equation" but which "belong" in the equation predicting program participation.

In the common effect model where $\epsilon=0$ or $U_1 - U_0 = 0$, (Case 1) **Heckman** (1978) shows that conventional instrumental variables methods would identify $\beta^* = \bar{\beta}$. One needs to find some variable or variables Z "uncorrelated with *U*" that affect "*D*" and are not already in the outcome equation (do not enter α and β or β^*, respectively).

In the model where ϵ is not a determinant of *D*, i.e., where

$$\Pr(D=1 \mid X,Z,\epsilon) = \Pr(D=1 \mid X,Z),$$

or

$$\Pr(D=1 \mid X,Z,Y_1) = \Pr(D=1 \mid X,Z),$$

the same conditions have to be satisfied for Z, i.e., "uncorrelated with *U*" and "not in the outcome equation." We can ignore the component $D\epsilon$ because

$$E(D\epsilon \mid X,Z) = E(\epsilon \mid X,Z,D=1)\Pr(D=1 \mid X,Z) = 0$$

since

$$E(\epsilon \mid X,Z,D=1) = 0.$$

Thus, the two special cases where $\bar{\beta}=\beta^*$ are cases where we can use conventional textbook instrumental variables models. (Heckman and Robb [1985, 1986] develop both of these cases.)

What about general cases? Consider equation (6') associated with β^*. Since the only source of dependence between the error term and D is through U and not $D(\epsilon-E(\epsilon \mid X,D=1))$ the instrumental variables method looks promising. If assumptions (C-1-b), (C-2), and (C-3-b) are satisfied, the method can be used to identify β^*. What is required is a variable Z that affects participation but does not enter the parameters in the equation of interest. This requires that

$$\begin{aligned} E(\bar{\beta}+\epsilon \mid X,Z,D=1) &= X(\beta_1 - \beta_0) + E(\epsilon \mid X,Z,D=1) \\ &= X(\beta_1 - \beta_0) + E(U_1 - U_0 \mid X,Z,D=1) \end{aligned}$$

does *not depend on Z.*

This is mechanically satisfied in the first *two* cases where $\epsilon = 0$ (Case 1), or $E(\epsilon \mid X,Z,D=1)=0$ (ϵ cannot be forecast by Z and X) (Case 2).[3] Even with only partial information about $U_1 - U_0$, standard economic models of selection into programs produce a violation of this condition. If individuals select into the program on the basis of the gain in unobservables or on the basis of the variables that are (stochastically) dependent on the gain in unobservables, this condition will not be satisfied. (See Heckman and Robb, 1985, 1986.)

Any application of the method of instrumental variables for estimating the mean effect of "treatment on the treated" in the case where the response to "treatment" varies among persons requires that a behavioral assumption be made about how persons make their decisions about program participation. The issue cannot be settled by a statistical analysis.

Consider an example that is widely cited as a triumph of the method of instrumental variables. Draft lottery numbers are alleged to be ideal instrumental variables for identifying the effect of military service on earnings (Angrist, 1990). The 1969 lottery randomly assigned different priority numbers to persons of different birth dates. The higher the number, the less likely was a person to be drafted. Persons with high numbers were virtually certain to be able to escape the draft. Letting "1" denote military service and "0" civilian service, if persons partly anticipate gain, $U_1 - U_0 = \epsilon$, or base their decisions to go into the military on variables correlated with unobservable components ϵ, persons with high Z for whom $D = 1$ (they serve in the military) are likely to have high values of ϵ. This violates assumption (C-3-b), and makes the birth date number an invalid instrument. It is plausible that the persons who are deciding to go into the military have more information at their disposal than analysts using standard data sets. If this information is at all useful in predicting the gain from going into the military, the draft number is not a valid instrument.

---

[3] Actually it is also satisfied in a third case where ϵ cannot be forecast by Z but can be forecast by X. But in general, this condition is very difficult to satisfy.

The draft lottery number is a poor instrument for another reason. Switching from a regime of a capricious draft to a lottery reduces uncertainty and is likely to change the investment behavior of persons of all levels of Z. In this instance, the switch from a draft to a lottery affects both β^* and $\bar{\beta}$ since it fundamentally alters schooling and job training investment decisions. Thus, knowing how military service affects earnings during the period of a lottery would not be informative on how military service affected earnings during the period of an ordinary draft.[4]

As a second example, it is sometimes suggested that cross-state variation in welfare benefit can be used as instrumental variables for estimating the effect of "treatment on the treated" for participants in training programs. Suppose that Y_0 refers to the earnings of untrained low-skill persons. Y_1 is earnings in the trained state. Parameters of welfare benefit determinants, Z, do not plausibly enter $\mu_0(X)$ or $\mu_1(X)$. But they could enter

$$E(\epsilon \mid X,Z,D=1) = E(U_1 - U_0 \mid X,Z,D=1)$$

because of the distribution of (U_1, U_0) is the same across the states, and if more generous welfare schemes discourage participation in the training program because they induce people to stay out of the market, then higher values of $U_1 - U_0$ would tend to be found *among program participants* in high-benefit states if program participants enter the program with at least partial knowledge of $U_1 - U_0$. Then assumption of (C-3-b) is violated, and cross-State variation benefits do not identify β^* through the method of instrumental values.'

---

[4] Yet another reason why the draft lottery is a poor instrument is the following: Z is likely to be an X. Persons with high Z (low chances of being drafted) are likely to be more attractive to employers investing in their workers. A person unlikely to be drafted is likely to be a better investment because he is less likely to be removed from the firm to perform military service. This causes (C-3-b) to be violated because Z is really an X.

5. Extension to the "Local Average Treatment Effect"

Imbens and **Angrist** (1994) distinguish between parameter@*-the effect of "treatment on the treated" also known as the effect of participation on participants-and the effect of treatment on those who change state in response to a change in Z. They define the latter as "LATE" for the Local Average Treatment Effect. This distinction is intended to allow for the possibility that individuals with distinct values of Z may have the same probability of participating in the program. For such people, differences in Z do not trigger changes in the probability of program participation. This distinction is implicit in (9), which is part of their definition of LATE when the denominator of that expression does not equal zero. If the denominator equals zero, (9) would not be defined. Obviously if there

---

It might be thought that since

$$E[U_1 - U_0 - E(U_1 U_0 \mid X, \mathrm{D} = 1) \mid x, \mathrm{D} = 1] = 0,$$

or equivalently that

$$E(\epsilon - E(\epsilon \mid X, D = 1) \mid X, D = 1) = 0$$

it follows that $E[U_1 - U_0 - E(U_1 - U_0 \mid \mathrm{X}, D = 1) \mid \mathrm{X}, Z, D = 1] = 0$. This is not true. In general

$$E(U_1 - U_0 \mid X, Z, D = 1) \neq E(U_1 - U_0 \mid X, D = 1).$$

Conditioning more finely on X and Z does not produce the same result as conditioning on X. Then even if

$$E(U \mid X, Z) = 0,$$

it does not follow that (C-l-b) is satisfied. In this case, Z is not a good instrument for identifying β^*. For similar reasons, (C-l-a) is unlikely to be satisfied even if

$$E(U \mid X,Z) = 0$$

because

$$E(D\epsilon \mid X,Z) = E(U_1 - U_2 \mid X,Z,D=1)Pr(D=1 \mid X,Z)$$

does not equal zero. In this case, Z is not a good instrument for identifying $\bar{\beta}$ either.

is no variation in the probability that $D = 1$ when Z_1 changes to Z_2 it is not possible to estimate β^*. Variation in Z is needed to estimate the parameters of interest.

Observe that if Z_1 and Z_2 are arbitrarily close to each other, and both numerator and denominator terms are continuously differentiable in Z, (9) can be defined in terms of derivatives:

$$\beta^* = \left.\frac{\partial E(Y \mid X, Z / \partial Z}{\partial \Pr(D = 1 \mid Z) / \partial Z}\right|_{Z = Z_1}$$

if the denominator is zero, the parameter is not defined.

Imbens and **Angrist** embellish the trivial requirement of a nonvanishing denominator with monotonicity conditions that ensure that the *numerator* of (9) is not zero. This requirement is inessential if the purpose of the analysis is to estimate how a change in Z at a certain level of X changes overall outcomes. If the goal is to estimate the change from "0" to "1" or from "1" to "0," their monotonicity requirement guarantees that one or the other of those changes is estimated by a change in the value of Z from Z_1 to Z_2.

The analysis of the preceding section clearly applies to their parameter. Indeed, the parameter defined by equation (9) is LATE if the additional (and generally irrelevant) monotonicity conditions for the numerator of the expression are satisfied. As before, instrumentalvariables define interesting economic parameters only if none of the unobserved components of the outcome equation are anticipated by the persons participating in the program.

6. Summary of Section II

Statistical assumptions often made in evaluation research are based on strong behavioral assumptions. This paper exposits how the widely used method of instrumental variables is based on the assumption (a) that persons respond identically to treatment or the assumption (b) that if responses are heterogeneous, persons do not make their decisions to participate in the program based on unobserved or forecastable components of program gains. Assumption (b) requires a strong form of ignorance about unobserved components of gain on the part of the people we study. It also implies that persons do not have private information that is useful in forecasting gains that they use in making their decisions but that are not available to the analyst. If these assumptions are incorrect, the method of instrumental variables is inconsistent when estimating the effect of treatment on the treated.

More general methods that solve the evaluation problem under less restrictive assumptions are presented in Bjorklund and Moffitt (1987), Heckman and Robb (1985, 1986), and Heckman and Smith (1995).

III. Randomized Trials

1. Introduction

This section of the paper discusses how randomized social experiments operate as an instrumental variable. For two types of randomization schemes, the fundamental experimental estimation equations are derived from the principle that experiments equate bias in control and experimental samples. Using conventional econometric representations, I derive the orthogonality conditions for the fundamental estimation equations. Randomization is a multiple instrumental variable in the sense that one randomization defines the parameter of interest expressed as a function of multiple endogenous variables in the conventional usage of that term. It orthogonalizes the treatment variable simultaneously with respect to the other regressors in the model and the disturbance term for the conditional population. However, conventional "structural" parameters are not in general identified by the two types of randomization schemes widely used in practice.

Randomized social experiments are now coming into widespread use. Their limitations and benefits are also beginning to be understood. Papers by Burtless (1995), Heckman (1992), Heckman and Smith (1995), and Moffitt (1992), among others, clarify the behavioral and statistical assumptions underlying the experimental method.

This section contributes to this literature and considers the social experiment as an instrumental variable. It develops the point that under the assumptions that justify its application, widely-used randomization schemes do not achieve their results by producing **exogeneity** of the treatment with respect to the population error term, as that term is ordinarily used in econometrics. Rather, these randomizations operate by *balancing* or equating the bias in the sample of persons randomized into a program with the bias in the sample of persons randomized out of the program. Randomization creates independence of the treatment effect with respect to other regressors and with respect to the error term in conditional populations. One randomization generates a multiple instrumental variable. Treatment effects as functions of an arbitrarily large number of endogenous variables can be identified from one randomization.

I develop these points for two distinct economic models: (a) a common effect model (treatment has the same effect on everyone with the same observed X characteristics) and (b) a variable effect model (treatment has different effects on everyone with the same observed X characteristics). The latter model is also known as a random effects model. The former is the one most widely used in applied work. **Heckman** and Robb (1985) and **Heckman** (1992) demonstrate the value in distinguishing between these two models in devising strategies for evaluating social programs.

I first consider randomization administered at the stage where persons apply to and are accepted into a social program and are then randomized out of the program. Randomization administered at that stage is widely used. Under the conditions specified in **Heckman** (1992) and **Heckman** and Smith (1995), this randomization identifies the mean gain to participating in the program for those who would usually participate in it. This mean gain is sometimes called the effect of treatment on the treated. I also consider samples produced by randomizing eligibility for the program. Before doing this, I briefly state the evaluation problem.

2. Randomization Balances Bias

First consider randomization at the stage where persons apply to and are accepted into a social program. Nowhere is it assumed that $E(U_1 \mid X) = 0$ or $E(U_0 \mid X) = 0$. Thus, X can fail to be exogenous in the conventional sense of that term. Yet randomized trials that do not disrupt the program, and are not subject to attrition or noncompliance, produce the data that can be used to consistently estimate the parameter (2) defined in Section II (the impact of treatment on the treated). This process highlights both the unusual nature of that parameter and the benefits of randomization.

To establish how randomization identifies **(2)**, it is instructive to introduce new variables denoted by *. $D^*=1$ denotes the event: "in the presence of randomization a person would have participated in the program except possibly for being randomized out." We may also define Y_1^* and Y_0^* to be the outcomes observed under a regime of randomization. Absence of randomization bias for mean gain is defined as

$$\text{(A-1)} \qquad E(Y_1 \mid D=1,X) = E(Y_1^* \mid D^*=1,X)$$

and

$$E(Y_0 \mid D=1,X) = E(Y_0^* \mid D^*=1,X)$$

where $Y_1^*=Y_1$ and $Y_0^*=Y_0$, i.e., randomization does not alter the outcome mean gain for the program being evaluated for all values of X.

Randomization operates *conditionally* on $D^*=1$ This is appropriate because parameter (2) in Section II is defined conditionally. In the subpopulation for which $D^*=1$ randomization operates by selecting persons into the program by a random device. $R = 1$ if a person is randomized into the program; $R = 0$ if a person is randomized out of the program. I assume that if $R = 1$, persons accept admission into the program and if $R = 0$, they do not obtain program services.

As a consequence of assumption (A-1), and the additional assumption that equates $R = 1$ or $R = 0$ with receipt of program services, using the definition $Y=Y_1D+Y_0(1-D)$

$$\text{1(a)} \qquad E(Y \mid D^*=1,R=1,X) = E(Y_1 \mid D=1,X) = g_1(X) + E(U_1 \mid D=1,X)$$

$$\text{1(b)} \qquad E(Y \mid D^*=1,R=0,X) = E(Y_0 \mid D=1,X) = g_0(X) + E(U_0 \mid D=1,X)$$

Randomization creates data that can be used to estimate counterfactual 1(b). Conditional mean 1(a) can be consistently estimated using ordinary observational data. If there is no randomization bias, and R is synonymous with receipt of services, both experiments and observational data are equally informative about I(a). I assume no randomization bias and for notational simplicity henceforth equate D and D^*, Y_1 and Y_1^*, and Y_0 and Y_0^*.

Subtract I(b) from I(a) to obtain

$$(2) \qquad E(Y_0 \mid D=1,R=1,X) - E(Y_1 \mid D=1,R=0,X) = \\ g_1(X) - g_0(X) + E(U_1 - U_0 \mid D=1,X) = E(\Delta \mid D=1,X).$$

This can be consistently estimated using sample counterparts to population means. If some of the Xvariables are continuous, a nonparametric kernel estimator for pointwise means can be constructed using conventional methods. (See, e.g., **Härdle**, 1990.) Nowhere is it necessary to assume that

$$(3) \qquad E(U_1 \mid X) = 0 \quad \text{or} \quad E(U_0 \mid X) = 0 \; .$$

In fact, it is clear from the definition of A that, in general,

$$(4) \qquad E(U_1 - U_0 \mid D=1,X) \neq 0.$$

To place randomization into a more familiar-looking instrumental variables framework, define $\tilde{U}_1$ as U_1 conditional on $D = 1$ and X and define $\tilde{U}_0$ as U_0 conditional on $D = 1$ and X. Then Y conditional on $D = 1$ and X may be written as $\tilde{Y}$, and

$$(5) \qquad \tilde{Y} = g_0(X) + [g_1(X) - g_0(X)]R + \{\tilde{U}_1 - \tilde{U}_0\}R + \tilde{U}_0$$

In this notation, it is not assumed that $E(\tilde{U}_0 \mid D=1,X) = 0$ nor is it assumed that $E(\tilde{U}_1 \mid D=1,X) = 0.$

Using definition (2) in Section II, and defining the mean-adjusted errors

$$\tilde{U}_0^* = \tilde{U}_0 - E(U_0 \mid D=1,X) \text{ and } \tilde{U}_1^* = \tilde{U}_1 - E(U_0 \mid D=1,X)$$

and $\underline{\varphi}(X) = g_0(X) + E(U_0 \mid D=1,X)$ we may rewrite equation (5) as

$$\tilde{Y} = \underline{\varphi}(X) + E(A \mid X,D=1)R + \tilde{U}_0^* + (\tilde{U}_1^* - \tilde{U}_0^*)R. \tag{6}$$

Conditioning on X, $\underline{\varphi}(X)$ is an intercept and (6) is a simple univariate regression defined for each value of X, given $D = 1$. Randomization makes R independent of $(\underline{\varphi}(X), \tilde{U}_0^*, \tilde{U}_1^*)$ conditional on $D = 1$ and X. The orthogonality conditions produced by randomization are

$$E[\tilde{U}_0^* + (\tilde{U}_1^* - \tilde{U}_0^*)R \mid R] = 0 \tag{7a}$$

and

$$E[\tilde{U}_0^* + (\tilde{U}_1^* - \tilde{U}_0^*)R] = 0 \tag{7b}$$

For all X given $D = 1$, (7a) identifies $E(A \mid X,D=1)$ and (7b) identifies $g_0(X) + E(U_0 \mid D=1,X)$ but not its individual components. Randomization makes R orthogonal to $\tilde{U}_0$, $\tilde{U}_1 - \tilde{U}_0$, $g_1(X)$, and $g_0(X)$. It does not make $g_0(X)$ or $g_1(X)$ orthogonal to U_0 or $U_1 - U_0$

Thus experiments do not in general identify $g_1(X)$ since in general $E(\tilde{U}_0 \mid X,D=1) \neq 0$. From this it follows that experiments of the type discussed in this section do not in general identify the structural parameters of the original equation but they identify parameter (2) (in Section II) provided that (A-I) is valid and persons assigned to treatment receive it and persons denied treatment do not. This point is obvious after it is recognized that randomization at the stage where persons have applied to and been.accepted into a program generates samples conditional on variables that are, in general, endogenous, in the conventional usage of that term.

These expressions simplify when there is a common effect model. In that case, $U_1 = U_0$ and then $\tilde{U}_0 = \tilde{U}_1 \overset{\text{def}}{=} \tilde{U}$. Then

$$E(\Delta \mid X,D=1) = g_1(X) - g_0(X) \tag{8}$$

and equation (5) may be written as

$$(9) \qquad \tilde{Y} = [g_0(X) + E(\tilde{U} \mid D=1,X)] + E(A \mid X,D=1)R + [\tilde{U} - E(\tilde{U} \mid D=1,X)].$$

Letting $\tilde{U}^{*} = \tilde{U} - E(\tilde{U} \mid D=1,X)$, orthogonality condition (9) becomes

$$E(\tilde{U}^{*} \mid R) = 0.$$

Again notice that in general $g_0(X)$ cannot be separated from $E(U \mid D=1,X)$.

A familiar form of the common effect model writes

$$g_0(X) = X\beta_0$$
$$g_1(X) = X\beta_1$$

so in the common effect model

$$E(A \mid D=1,X) = X(\beta_1 - \beta_0)$$

where the conditioning on $D = 1$ is often left implicit. An even more familiar form of the common effect model writes

$$g_0(X) = X\beta_0$$
$$g_1(X) = X\beta_0 + \alpha$$

This is the dummy endogenous variable model (Heckman, 1978). In this case

$$(10) \qquad \tilde{Y} = X\beta_0 + R\alpha + \tilde{U}.$$

Randomization ensures that R is independent of both $\tilde{U}$ and X. It does not ensure that X is independent of $\tilde{U}$. The orthogonality between R and X induced by an experiment implies that any dependence between X and $\tilde{U}$ does not affect the identifiability of α. Randomization creates an orthogonal regressor model for the subpopulation defined conditional on $D = 1$. Since R is independent of X and $\tilde{U}$, α is identified even if X is not orthogonal to $\tilde{U}$, and β_0 is not identified.

Observe that one randomization identifies an entire function $E(A \mid X, D=1)$ over the support of X (i.e., the values of X where this parameter is defined). In principle, $E(A \mid X, D=1)$ can be an infinite-dimensional function of X . Hence, in this sense randomization is a multiple instrumental variable.

Observe further that randomization enriches the support of X in the following way. Suppose in the population that

$$\text{Support } (X \mid D=1) \neq \text{Support } (X \mid D=0).$$

Then in the subset of X values for which there is no overlap, observational methods cannot obtain comparisons for all X values and $E(A \mid X, D=1)$ cannot be identified for all X.[6] Randomization creates a balanced support set because

$$\text{Support } (X \mid D=1, R=1) = \text{Support } (X \mid D=1, R=0)$$

Unless

$$\text{Support } (X \mid D=1) \subset \text{Support } (X \mid D=0)$$

randomization enlarges the support set over which $E(A \mid X, D=1)$ can be defined and estimated. However, unless $0 < \Pr(D=1 \mid X) < 1$ randomization does not identify $E(A \mid X, D=1)$ for all possible values of X. (See, e.g., Rosenbaum and **Rubin**, 1983.) An extreme example of the benefit of randomization in enlarging the support set occurs when for certain values of X, the event $D = 0$ does not occur, i.e., $\Pr(D=0 \mid X) = 0$.

But $D = 1$ occurs with positive probability for all values of X:

---

[6] **Heckman** and Roselius (1994); **Heckman**, Ichimura, Smith, and Todd **(1995a,b)**; **Heckman**, Ichimura, and Todd (1995a,b); and **Heckman**, Ichimura, and Todd (1995, revised 1996) document that failure of a common support condition is a major component of what is traditionally regarded as selection bias.

$$0 \ c \ \Pr(D=1 \mid X) \ c \ 1.$$

In this case, randomization expands the support of X given $D = 1$ to the entire support of X. It permits identification of $E(A \mid X, D=1)$ for all possible values of X.

Observe that, in general, experiments defined conditional on $D = 1$, do not identify $E(A \mid X)$, the effect of selecting a person at random from state "0" and moving the person to "1." However, if the common effect model is assumed, $U_1 - U_0 = 0$ and experiments conducted on populations defined conditional on $D = 1$ recover $E(A \mid X)$, the effect of selecting a person from the general population and placing him/her in the program. For in that case, $E(A \mid X) = E(A \mid D=1,X)$. (See **Heckman**, 1992, or **Heckman** and Smith, 1995.)

In one case where responses to treatment are heterogeneous, randomization is administered to those for whom D would have been "1" in the absence of randomization and the parameter (3) is nonetheless identified as follows. Suppose at the time agents enroll in the program they forecast *their* gain to be the total population mean gain. Then clearly (2) equals (3) and the experiment identifies both parameters (**Heckman** and **Robb**, 1985). However, in general, if $U_0 \neq U_1$, $E(A \mid X) \neq E(A \mid D=1,X)$.

4. Randomization of Eligibility

Randomization of eligibility for a program is sometimes proposed as a less disruptive alternative to randomization of admission among accepted applicants (Heckman, 1992; **Heckman** and Smith, 1993; **Angrist** and Imbens, 1991). In this section, I show that this type of randomization can be placed in an instrumental variable framework. Consider a population of persons ordinarily eligible for a program. For simplicity, this conditioning is kept implicit. Let e = 1 if a person is kept eligible after randomization; e = 0 if the person loses eligibility. Assume that assignment to eligibility does not disturb the underlying stochastic structure and that it is independent with respect to the outcome measures:

$$\text{(A-2)} \qquad (Y_0, Y_1, D, X,) \perp\!\!\!\perp e.$$

Assuming $P(D=1 \mid X \neq 0)$,

$$(11) \qquad \frac{E(Y \mid e = 1, X) - E(Y \mid e = 0, X)}{P(D = 1 \mid X)} = E(\Delta \mid D = 1, X).$$

To prove this, use the law of iterated expectations to obtain

$$12(a) \quad E(Y) e=1,X) =$$
$$E(Y_1 \mid D=1,e=1,X)P(D=1 \mid e=1,X) + E(Y_0 \mid D=0,e=1,X)P(D=0 \mid e=1,X)$$

and

$$12(b) \quad E(Y \mid e=0,X) =$$
$$E(Y_0 \mid D=1,e=0,X)P(D=1 \mid e=0,X) + E(Y_0 \mid D=0,e=0,X)P(D=0 \mid e=0,X)$$

From (A-2)

$$P(D=1 \mid e=1,X) = P(D=1 \mid e=0,X) = P(D=1 \mid X)$$

so that the result follows by subtracting 12(b) from 12(a) and dividing by $P(D=1 \| X)$ provided $P(D=1 \mid X) \neq 0$. Replacing population moments by sample moments, (11) is a version of what is sometimes called Bloom's estimator for attrition from a program. (See Angrist and Imbens, 1991; and Heckman, Smith, and Taber, 1994, revised 1995, for a discussion of this estimator.)

I now present an instrumental variables interpretation of this estimator. Using the law of iterated expectations and (A-2), and the notation introduced in Section 2

$$(13) \qquad Y = \varphi(X) + [E(\Delta \mid D=1,X)]P(D=1 \mid X)e + \tilde{U}_0^* + (\tilde{U}_1^* - \tilde{U}_0^*)P(D=1 \mid X)e$$

where

$$\tilde{U}_0^* = \tilde{U}_0 - E(\tilde{U}_0 \mid D=1,X)$$
$$\tilde{U}_1^* = \tilde{U}_1 - E(\tilde{U}_1 \mid D=1,X)$$

and

$$\underline{\varphi}(X) = g_1(X) + E(U_0 \mid D = 1, X)$$

Observe that by random assignment of e,

$$e \perp\!\!\!\perp \{U_0^* + e[(U_1^* - U_0^*)]P(D = 1 \mid X)e\}$$

so that orthogonality (really independence) is an immediate consequence of the randomization. Again, there is no requirement that X be independent or orthogonal with respect to U_1, or U_0. Hence, under standard rank conditions one can consistently estimate $\mathbf{E}(\mathbf{A} \mid D=1,X)P(D=1 \mid X)$. Assuming that one can consistently estimate $P(D=1 \mid X)$ one can estimate $\mathbf{E}(\mathbf{A} \mid D=1,X)$ the effect of treatment on the treated, by dividing the instrumental variables estimator of the product of the two terms by $P(D=1 \mid X)$. Note that this randomization identifies $\mathbf{E}(\mathbf{A} \mid D=1,X)$ but not $\mathbf{E}(\mathbf{A} \mid X)$, except in the special cases where the two parameters are the same.

5. Concluding Remarks on Section III

This section considers randomization as an instrumental variable. Two types of randomizations are considered: (a) randomization of eligibility for a program and (b) randomization of admission into the program among eligible persons who would ordinarily be admitted into the program. The second type of randomization is widely used in conducting social experiments. Using a conventional separable-in-the-errors representation of equations, we have shown the orthogonality conditions that are produced by the two types of randomization schemes and how they identify a central parameter in program evaluation studies-the effect of treatment on the treated. One randomization serves to identity this parameter as a function of multiple endogenous variables as conventionally defined in econometrics.

Balancing the bias in experimental and control samples is the fundamental source of identification from experiments. Such balancing in no way depends on separability of errors from equations as conventionally assumed in econometrics nor does it require that the X be either independent or orthogonal with respect to the U. The method of moments analogs to (2) or (10) can be implemented nonparametrically. The balancing conditions are the basic estimating equations for experiments from which the orthogonality conditions of this paper have been derived.

The fact that parameter (2) defined in Section II is not conventional has been the source of some confusion. It combines both structural portions with conditional means of the errors (the U_1 and U_0). Experiments conducted at a stage where persons would ordinarily enter a program are not designed to consistently estimate the g_0 and g_1 functions and in general they do not. Experiments make the treatment variable orthogonal to the error and the other regressors thus separating the estimation of treatment effects from the estimation of the other parameters of the model.

Only under special conditions does either type of randomization discussed in this paper identify parameter (3)-the effect of moving a randomly selected person in the general population from state "0" to state "1." This is an intrinsically more difficult parameter to estimate using social experiments because in most societies people cannot be forced to participate in programs against their will. It is more difficult to estimate $E(Y_1 \mid D=0,X)$ than $E(Y_0 \mid D=0,X)$ if responses to treatments are heterogeneous, and are partly anticipated at the time decisions to enroll in the program are made. By the same token, if there is attrition from the program, it may also be difficult to estimate (2) except under special conditions discussed in Bloom (1984), Heckman, Smith, and Taber (1997), or Hotz and Sanders (1994).

Other types of randomization might be used besides the two types considered in this paper. For example, if interest centers on estimating $Y = g(X) + U$ and X is not independent of U, X might be experimentally varied as in the negative income tax experiments or in the electricity experiments. In this case, experimental variation in X can be used in the conventional way to produce an instrument for an endogenous variable. See the discussion in Heckman (1992).

IV. Matching as an Evaluation Estimator

1. Introduction

Using the methods developed in previous papers, Heckman, Ichimura, and Todd (1995a,b), I summarize the evidence on the empirical performance of matching as an estimator of program outcomes for the JTPA program. JTPA is the largest federally funded training program for disadvantaged workers in the United States. In previous joint research (Heckman, Ichimura, Smith, and Todd, 1995a,b), we report evidence that suggests that a form of kernel matching plus regression appears to be an appropriate method for evaluating the JTPA program provided that one is interested in certain sample averages.

The JTPA comparison group data differ from the data usually used in nonexperimental studies in that comparison group members share more features in common with program applicants. They are eligible for the program, reside in the same local labor markets, and complete the same survey instrument. Previous nonexperimental studies often had insufficient geographical identifier information and were not able to match comparison group members to the same labor market as participants, had insufficient information to determine eligibility status, and relied on earnings and other data collected from different survey questionnaires. To determine how important the improved properties of our comparison group are in obtaining reliable estimates of program impacts, we also analyze comparison group samples drawn from the Survey of Income and Program Participation (SIPP) data. The SIPP samples are not geographically matched and are collected using a different survey instrument. The SIPP data have enough information to determine program eligibility status, which makes it an attractive source of comparison group data.

For all major demographic groups, we find that kernel-based matching works well estimating certain average effects in large samples (with 500 or more persons in comparison and treatment groups) provided that analysts have access to information sufficiently rich to determine the probability that persons participate in the program and provided that participants and comparison group members are matched in the same labor markets and are interviewed with the same questionnaires. A hybrid method developed in Heckman, Ichimura, and Todd (1997a) that combines kernel-based matching derived from local-linear regression methods generally works even better.

Our empirical evidence suggests that a major source of bias in previous evaluations of training programs comes from the relatively crude data used in those studies. As noted above, randomized trials accomplish many things at the same time: (1) they place persons in the same labor market, (2) they eliminate selective differences between treatments and controls through the random selection mechanism, (3) they usually administer the same questionnaire to treatments and controls, and (4) they balance the distribution of characteristics (the X) between participants and nonparticipants. The first, third, and fourth features of randomized data could also be incorporated into nonexperimental studies, though in the past they have not been. One goal of my joint research is to determine the relative importance of these sources to the evaluation bias that arises from using nonparticipants to proxy the outcomes of randomized-out participants.

Previous nonexperimental research used very crude data. Table 1 summarizes the quality of data available in several influential studies of training. Variables were measured only on an annual basis, not on the monthly or quarterly basis used in this study. Comparison group members did not reside in the same labor markets as treatment group members. Insufficient information existed to determine whether or not comparison group members were eligible for the program. Different survey instruments were used to collect data on treatment and control group members. No information on short-term labor force dynamics was available. Under these severe data limitations, it is not surprising that nonexperimental estimators failed to produce estimates that are consistent with the experimental ones.

The wide range of estimates produced in these studies led to skepticism concerning the reliability of the statistical methods used (LaLonde, 1986). My joint papers with Ichimura, Smith and Todd show that different statistical methods vary widely in their effectiveness, but the quality of the data used to perform the analysis also plays an important role. Placing eligible nonparticipants in the same labor market and using the same survey instrument produces a comparison group that is extremely close to an experimental control group in measuring the impact of the program. If, in addition, rich data on individual characteristics-including labor market histories-are available, then a nonexperimental matching estimator can be successfully implemented to estimate the impact of the program. By considering various matching estimators applied to different data sets, we determine which of the estimators is most reliable. We also demonstrate the benefit of having available a rich set of conditioning variables.

The balance of this paper is organized in the following way. Section 2 briefly describes the JTPA data we use and summarizes the evidence on the main determinants of participation in the JTPA program, presented in Heckman and Smith (1994).

Section 3 compares the performance of alternative matching estimators. We measure the effectiveness of a nonexperimental estimator by how well it eliminates differences between the

Table 1

Comparison Groups Used in Different Studies*

CLMS-Based Studies

Study	Ashenfelter (1978)	Ashenfelter and Card (1985)	Dickinson, Johnson and West (1987)	Westat (Rupp and Bryant) (1986)
Program, Year, Outcomes	MDTA classroom trainees, first 3 months of 1964, 1965-1969 annual social security record earnings	CETA, 1976, 1977-1978 annual social security record earnings	CETA, 1976, 1978 annual social security record earnings	CETA, 2 cohorts, 1977-1978 annual social security record earnings
(1) Comparison group in the same labor market?	NO	No	NO	NO
(2) Same questionnaire administered to comparison and treatment groups?	Yes	Yes	Yes	Yes
(3) Matching criteria (criteria for membership ion comparison sample, also called "screening" criteria)	None Specified	(a) 1975 earnings <=$20K Household income< = $30K (b) In labor force, March 1976 Matched on age (persons > =21 used)	(Matching based on a metric over vectors of variables) Matched on predictors of 1978 earnings including lagged earnings (1975-1970), unemployment in 1975, worked in public sector, sex, and demographics. In labor force, March 1976	Match on 1976 earnings, change in 1976 earnings (1975-1976), 1974-1975 change in earnings, demographics, 1975 labor force status, family income (for 1976-1977 cohort one year previous for 1975-1976 cohort). Either in labor force, 1975, or at interview March 1976. Three matching groups based on income.
(4) Eligibility to program known for comparison group members?	NO	NO	No	NO
Variables used in analysis				
Age, Race, Sex	Yes (NO age restriction)	Yes (Age>= 21 years)	Yes (Age 21-65)	Yes (Age 14-60)
Education	NO	Yes	Yes	Yes
Training History	No	No	No	No
Children	NO	NO	NO	No
Employment Histories	NO	NO	Yes (recent)	Yes (recent)
Hours Worked	NO	Yes	Yes	Yes

BEST COPY AVAILABLE

Unemployment Histories	No	No	Yes (recent)	Yes (recent)
On Welfare	No	No	Yes	No
Earnings Histories"	(Annual earnings) 5 years pre-program 5 years post-program	(Annual earnings) 2 years pre-program 2 years post-program	(Annual earnings) 2 years pre-program 2 years post-program	(Annual earnings) 4 years pre-enrollment

Other Studies

	National Supported Work		JTPA Data
Study	**LaLonde** (1986)	Fraker and Maynard (1987) and **LaLonde** and Maynard (1986)	**JTPA** Data
Program, Year, Outcome Variable	Annual earnings 1978 annual social security earnings and PSID earnings	1977, 1978, 1979 annual earnings for AFDC recipients and for youth	Quarterly and Monthly earnings 1987-1989
(1) Comparison group in the same labor market?	No	No	Yes
(2) Same questionnaire administered to comparison and treatment groups?			
(3) Matching criteria (criteria for membership ion comparison sample, also called "screening" criteria)	PSID: men and women who are household heads 19751979 CPS: matches March 1976 CPS earnings with SSA earnings. Persons with 1976 income **< =20K** and household income **< =30K**	Three samples: **I. Eligible in sample period:** For youth: high school dropout -- exclude in-school youth. For AFDC: Age of youngest child, receipt of welfare matching **II. Cell Matching** Based on predictors of 1979 SSA earnings of eligibles: earnings prior to program participation, demographics, education, family income, changes in earnings **III. Stratified matches on imputed 1979** earnings earnings estimated on eligible nonparticipant sample plus demographic criteria (race, sex). Same criteria for prediction as in II.	Persons screened to be eligible for **JTPA**; out-of-school youth, no disabled persons; Title II-A only.
(4) Eligibility for program known for comparison group members?	No	No	Yes

BEST COPY AVAILABLE

Variables used in analysis			
Age, Race, Sex	Yes: Women AFDC recipients 20-55, Males <=55	Yes: Women AFDC recipients 20-55, Males <=55	Yes
Education	Yes	Yes	Yes
Training History	No	No	Yes
Children	Yes	Yes	Yes
Employment Histories	No	No	Yes
Hours worked	No	No	Yes
Unemployment Histories	No	No	Yes
Welfare Receipt	Yes	Yes	Yes
Earnings Histories	**Two** years post-program Two years pre-program	Two years post-program **Two** years pre-program	Five years of pre-program earnings (monthly earnings)

* "Used" means either in outcome equations or matching equations.

** CLMS data matched social security longitudinal records to March CPS data for 1976 and 1977. The CPS data are for comparison group members. Only SSA data on longitudinal earnings are available for both groups. **All** of the personal and family information available in the CPS, including short-term employment and labor-force participation histories are available but not necessarily used in the analysis. The **CLMS** studies all use the social security earnings data.

BEST COPY AVAILABLE

nonexperimental comparison group and the randomized-out controlgroup. Matching estimators that smooth are more effective than simple nearest neighbor schemes often used in the literature. Estimators based on the Mahalanobis metric, recommended by Rubin (1979) and used in previous controversial evaluations of training programs (see, e.g., Bryant and Rupp, 1986, and Dickinson et al., 1986), have little theoretical justification and perform very badly.

Section 4 studies the effectiveness of different estimators when progressively coarser conditioning sets are used in estimation. Matching estimators are found to perform well for all demographic groups only when data on recent labor market histories is incorporated in estimating the propensity score. Section 5 analyzes comparison group samples drawn from the SIPP data to assess the importance of controlling for geographic location and of using the same survey instrument in collecting comparison group data. These results indicate that geographical proximity and uniformity of the instrument across treatment and comparison group samples are necessary for a successful matching scheme. This evidence confirms the importance of local labor markets in determining wage setting.

2. The JTPA Data Determinants of Program Participation

The JTPA experiment was commissioned in 1986 by the Department of Labor for the purpose of evaluating ongoing JTPA employment and training programs. These programs provide on-the-job training, job search assistance, and classroom training to disadvantaged youth and adults under Title IIA of the Job Training Partnership Act. The experiment gathered longitudinal data on a group of treatments, controls, and eligible nonparticipants (ENPs). The samples used in this study come from four of the sixteen JTPA training sites participating in the study.[7] In the experiment, two-thirds of the applicants were assigned to treatment and one-third were randomized out and denied access to JTPA services for eighteen months to form a control group. Random assignment covered some or all of the period from November 1987 to September 1989 at each site. A total of 20,601 persons participated in the experiment.

---

[7] Kemple, Doolittle, and Wallace (1993) provide detailed description of all sixteen experimental sites.

The ENP-comparison group sample is based on a sample of dwelling units drawn at the four sites.[8] Screening interviews were administered to a random sample of dwelling units in the included areas at each site. Attempts were then made to administer a survey to all persons identified in the screening surveys who were: (1) eligible for JTPA via economic disadvantage, (2) 16 to 54 years of age, (3) not in junior high or high school, and (4) not permanently disabled. This process resulted in a sample of 3,004 ENPs. Individuals in the adult samples are of ages 22 to 54 and those in the youth samples are of ages 16 to 21. The Long Baseline Survey (LBS) collected retrospective monthly data on demographic characteristics, earnings histories, labor market histories, participation in government transfer programs, and participation in schooling or training activities. A follow-up survey, administered twelve to twenty-four months after the LBS, collected similar information. The response rate for this survey was around 84 percent. The sample only includes persons who (1) had a follow-up interview scheduled at least eighteen months after random assignment, (2) responded to the survey, and (3) had usable earnings information for the eighteen months after random assignment. Smith (1994) contains additional information on the design and collection of the ENP sample.

Persons were assigned to the control group only after they had applied to the JTPA program, been declared eligible, and been accepted into the program. The ENPs and controls were administered the same survey instruments, which included detailed retrospective questions on labor force participation, job spells, earnings, marital status, training and schooling activities, transfer program participation, and other demographic characteristics.

We combined the information from the LBS and follow-up surveys for ENPs and for controls to form a thirty-six-month panel data set. It is divided into the eighteen-month period before random assignment (or before the time of eligibility determination for ENPs) and the eighteen-month post-random-assignment period. Additional details on these surveys and on the construction of the samples used in our analysis are given in Smith (1994).

---

[8] To reduce survey costs, the sampling frame excluded low poverty areas containing up to, but not more than, 5 percent of those in each site with incomes at or below 125 percent of the poverty level in 1980. In the remaining areas of each site, each dwelling unit had an equal probability of selection.

Heckman and Smith (1994) present an extensive analysis of the determinants of participation in the JTPA program. Because the propensity score (P(x)) is a main ingredient of our empirical approach, I briefly summarize their findings concerning the relative importance of background characteristics, recent labor force status, and earnings histories in the participation process conditional on having been determined eligible for the program.

The central conclusion of their work is that, for all demographic groups, recent unemployment histories are good predictors of participation in training programs. Trainees enter these programs as a form of job search. For adult women, recent marital histories are also important. Recently divorced or separated women are much more likely to participate in the program than others. Job training is for people in transition who are seeking work. Models based on variables that predict job-seeking are much better able to predict participation than are models that include only demographic characteristics and variables relating to earnings and employment dynamics-the data typically used in previous evaluations of training programs. None of the major studies listed in Table 1 had access to data on recent unemployment histories. A dip in earnings is observed for participants shortly before they apply to the program. This phenomenon was first noted by Ashenfelter (1978) for an earlier training program. It is most pronounced for adult men and least pronounced for female youth. For each group, there is a recovery from the drop in mean earnings over the post-program period to a level above the preprogram mean. The earnings of ENP members have a mild dip prior to enrollment in the program and generally exhibit much greater stability.[9]

Heckman and Smith (1994) document that Ashenfelter's dip is only part of the story. An important finding in their work is that unemployment peaks just prior to the date of enrollment for participants compared to no change in unemployment status for nonparticipants. Unemployment increases both because employed persons lose their jobs and because persons previously out of the labor force enter it. Participants have a greater attachment to the labor force than nonparticipants. They experience a greater number of labor force transitions (where a transition is defined as moving from the state of employment to unemployment or vice versa) than nonparticipants.

---

[9] Heckman and Smith (1994) demonstrate that program eligibility rules produce the dip in earnings observed for ENP members. The dip occurs in the middle of the six-month income eligibility interval, and not at the end of the interval where it occurs for participants.

The analysis presented in our companion paper is developed under sufficiently general conditions to allow for nonparametrically estimated propensity scores. A major disadvantage of using a fully nonparametric estimation method is that with many continuous variables rates of convergence tend to be slow. With discrete variables there can be problems with small cell sizes. A semiparametric method that allows for choice-based sampling, such as that proposed in Todd (1995), circumvents these difficulties and could be used to estimate the propensity scores. However, she shows that in practice parametric procedures can perform as well or better than semiparametric procedures even under misspecifications of the residual distribution. Therefore, in my joint empirical work, we follow the literature in statistics and estimate parametric logit choice models using the weighting procedure developed in Rao (1965) and applied in Manski and Lerman (1977) to account for choice-based sampling.

The variables used in the participation models are described in Table 2. Tables 3(a)-(d) give the estimated logit coefficients for each of the demographic groups. Regressors in the models were chosen to maximize the within-sample correct prediction rates that are shown in Table 4.

2. The Support Problem: Comparison of the Densities of the Estimated Propensity Scores

Figures l(a)-l(d) plot the kernel-smoothed density of the estimated propensity scores, P_i, for each of the demographic groups. In our samples, the Rosenbaum-Rubin condition required for strong ignorability (1985) is not satisfied. An important finding is that the empirical support of the estimated P_i for controls and ENPs does not overlap in some regions. The support estimated for participants covers a wider range of values than for ENPs. Propensity scores for ENPs are more closely concentrated at low values around zero.

The main implication of this finding is that nonparametric matching methods can only be meaningfully applied over the region of overlapping support, where matched earnings for each program applicant can be estimated reliably. Simple nearest neighbor matching estimators could be mechanically applied, but matches for controls with high P_i values are poor. In addition, if nearest neighbor matching is performed with replacement, impact estimates will be sensitive to the inclusion or omission of a few persons in the ENP samples who are used repeatedly as matches for high P_i

Table 2

Definition of Variables

Site Indicators: 1/0 variables indicating whether respondent lives in Fort Wayne, Jersey City or Corpus Christi. Providence is the omitted category.

Race/Ethnicity Indicators: I/O variables indicating whether respondent is black, Hispanic or in the other category which includes Asian and Native American. White is the omitted category.

Age: I/O variables indicating whether the age of the respondent falls in the categories 19-21, 30-39, 30-39, 50-54. For youth, 16-18 is the omitted category. For adults, 22-29 is the omitted category.

Highest Grade Completed Indicators: I/O variables indicating whether the education of the respondent falls within the following categories: less than tenth grade, tenth or eleventh grade, one to three years of college, and four or more years of college. Twelfth grade is the omitted category. For youth samples, there were very few observations in the four or more years of college category, so this category was combined with the one to three years of college category.

Marital Status Indicators: 1/0 variables indicating whether the respondent falls within the categories: single at the point of random assignment and never married, single at the time of random assignment or eligibility determination but married in the prior year, single at the time of random assignment or eligibility determination but last married over one year prior. The omitted category is married at the time of random assignment. For female youth, twenty-four months of marital status instead of one year is used in defining the variables.

Vocational Training and Classroom Training Indicators: I/O variables indicating whether the respondent was in vocational or classroom training at the time of random assignment or eligibility determination and whether the respondent received such training in the prior year. For adult men and women, the categories are currently in vocational training and ever had vocational training. For adult women, additional categories indicate whether the individual was in classroom training at the time of random assignment/eligibility determination and whether the individual was in schooling in the previous one to three months or the previous four to six months.

AFDC or Food Stamp Recipient Indicator: 1/0 variables indicating whether the respondent received AFDC and/or Food Stamps at the time of random assignment/eligibility determination. The categories are: food stamps only, AFDC only or both AFDC and food stamps. The omitted category is received neither.

Number of Job Spells Indicator: I/O variables indicating the number of job spells observed in the eighteen months prior to random assignment or eligibility determination. The categories are one job spell, two job spells, and three or more job spells. The omitted category is zero job spells.

Most Recent Labor Force Status Indicators: I/O variables indicating the respondent's most recent labor force statuses within the seven months up to and including the month of eligibility determination. The categories are: unemployed to employed, out of the labor force to employed, employed to unemployed, unemployed to unemployed, out of the labor force to unemployed, employed to out of the labor force, unemployed to out of the labor force, out of the labor force to out of the labor force. The omitted category is employed to employed.

Number of Labor Force Transitions in the Past 24 Months: 1/0 variables indicating how many times the individual changed from employed to unemployed or vice versa in the past two years. The categories are one transition, two transitions, and three or more transitions. The omitted category is zero transitions.

Table 3(a)
Estimated Coefficients from Weighted Logit*

Response Variable: D=1 if Control, 0 if ENP Sample of Adult Men, 536 controls and 423 ENPs				
Variable	Coeff	Standard Error	T-Statistic	Prob> | T |
Intercept	-6.76	-1.58	-4.28	0.0000
Fort Wayne, Indiana	2.44	0.93	2.63	0.0086
Jersey City, New Jersey	0.24	0.91	0.26	0.7946
Providence, Rhode Island	1.81	1.00	1.81	0.0703
Black	0.60	0.70	0.86	0.3919
Hispanic	0.60	0.91	0.65	0.5133
Race other than White, Black or Hispanic	1.14	1.33	0.85	0.3942
Age 30 to 39	-0.53	0.52	-1.02	0.3072
Age 40 to 49	-0.54	0.75	-0.72	0.4727
Age SO to 54	-0.18	1.45	-0.12	0.9016
Less than Tenth Grade	-0.54	0.68	-0.80	0.4245
Tenth-Eleventh Grade	0.58	0.65	0.90	0.3665
One to Three Years College	1.13	0.76	1.50	0.1338
Four or More Years College	-1.85	1.13	-1.64	0.1020
Last Married 1-12 Months Prior to Random Assignment	0.97	1.32	0.73	0.4636
Last married More than Twelve Months Prior to Random	0.42	0.90	0.46	0.6423
Assignment	0.94	0.62	1.51	0.1316
Single, Never married	0.13	0.72	0.18	0.8582
Children Less Than Six	1.87	0.86	2.18	0.0293
Unemployed to Employed	1.81	1.26	1.44	0.1496
Out of Labor Force to Employed	3.84	0.77	4.98	0.0000
Employed to Unemployed	4.27	0.91	4.69	0.0000
Unemployed to Unemployed	4.73	1.34	3.52	0.0004
Out of Labor Force to Unemployed	4.30	1.25	3.45	0.0006
Employed to Out of labor Force	3.68	1.88	1.96	0.0506
Unemployed to Out of Labor Force	1.50	1.16	1.30	0.1952
Out of Labor Force to Out of Labor Force	0.43	0.89	0.49	0.6269
One Job Spell in Eighteen Months Prior to Random Assignment	0.56	0.96	OS8	0.5615
Two Job Spells	1.59	1.06	1.51	0.1309
Three or More Job Spells	1.72	1.05	1.63	0.1025
Currently in Vocational Training	-0.27	0.18	-1.51	0.1298
Ever had Vocational Training	-0.27	0.18	-1.51	0.1298
Number of Household Members				

The logistic regression is weighted because our data are characterized by choice-based sampling. The weighting scheme is described in Manski and Lerman (1977).

BEST COPY AVAILABLE

Table 3(b)
Estimated Coefficients from Weighted Logit*

Response Variable: D= 1 if Control, 0 if ENP
Sample of Adult Women, 720 controls and 910 ENPs

Variable	Coeff	Standard Error	T-Statistic	Prob> | T |
Intercept	-6.18	0.36	-17.00	0.0000
Fort Wayne, Indiana	1.24	0.26	4.82	0.0000
Jersey City, New Jersey	0.99	0.25	4.03	0.0001
Providence, Rhode Island	0.64	0.28	2.26	0.0236
Black	0.10	0.21	0.48	0.6323
Hispanic	0.32	0.24	1.37	0.1696
Race other than White, Black or Hispanic	-0.06	0.40	-0.15	0.8790
Age 30 to 39	-0.2%	0.17	-1.64	0.1010
Age 40 to 49	-0.20	0.22	-0.91	0.3629
Age 50 to 54	-0.12	0.35	-0.35	0.7285
Less than Tenth Grade	-0.32	0.19	-1.67	0.0940
Tenth-Eleventh Grade	-0.15	0.20	-0.75	0.4513
One to Three Years College	0.01	0.25	0.04	0.9663
Four or More Years College	-0.51	0.48	-1.04	0.2971
Last Married 1-12 Months Prior to Random Assignment	1.50	0.34	4.44	0.0000
Last married More than Twelve Months Prior to Random Assignment	1.93	0.21	9.16	0.0000
Single, Never married	0.64	0.22	2.89	0.0038
Children Less Than Six	-0.24	0.16	-1.50	0.1331
Unemployed to Employed	1.11	0.35	3.19	0.0014
Out of Labor Force to Employed	0.24	0.41	0.57	0.5687
Employed to Unemployed	2.30	0.32	7.24	0.0000
Unemployed to Unemployed	2.33	0.34	6.93	0.0000
Out of Labor Force to Unemployed	2.19	0.38	5.77	0.0000
Employed to Out of labor Force	0.38	0.35	1.06	0.2878
Unemployed to Out of Labor Force	1.66	0.50	3.34	0.0008
Out of Labor Force to Out of Labor Force	0.75	0.28	2.65	0.0081
Not Employed in Last Twenty Four Months Prior to Random Assignment	0.00	0.31	0.01	0.9908
One labor Force transition in last Twenty four Months	0.76	0.26	2.86	0.0042
Two Labor Force Transitions	0.55	0.30	1.84	0.0659
Three or More Labor Force Transitions	1.25	0.31	3.97	0.0001
Currently in Vocational Training	2.30	0.36	6.46	0.0000
Ever in Vocational Training in Last year	0.04	0.19	0.23	0.8218
In Schooling at Random Assignment	-0.30	0.33	-0.92	0.3587
Last in Schooling one to Three Months Ago	1.09	0.35	3.09	0.0020
last in Schooling Four to Six Months Ago	0.66	0.73	0.91	0.3652

[*] The logistic regression is weighted because our data are characterized by choice-based sampling. The weightin scheme is described in Manski and Lerman (1977).

BEST COPY AVAILABLE

Table 3(c)
Estimated Coefficients from Weighted Logit*

Response Variable: D=1 if Control, 0 if ENP
Sample of Male Youths, 279 controls and 96 ENPs

Variable	Coeff	Standard Error	T-Statistic	Prob> | T |
Intercept	-5.17	0.63	-8.15	0.0000
Fort Wayne, Indiana	-1.30	0.66	-1.97	0.0483
Jersey City, New Jersey	-1.64	0.70	-2.36	0.0183
Providence, Rhode Island	0.80	0.43	1.88	0.0597
Black	0.50	0.55	0.91	0.3621
Hispanic	1.00	0.64	1.58	0.1139
Age 19 to 21	0.64	0.51	1.25	0.2095
Less than Tenth Grade	0.57	0.56	1.03	0.3047
Tenth-Eleventh Grade	0.69	0.46	1.49	0.1357
One to Three Years College	-1.43	0.99	-1.44	0.1485
Married and Living with Spouse	-0.18	0.73	-0.25	0.8023
Children Less than Six	-1.39	0.76	-1.84	0.0657
Unemployed to Employed	2.60	0.62	4.19	0.0000
Out of Labor Force to Employed	0.02	0.52	0.03	0.9740
Employed to Unemployed	2.46	0.66	3.73	0.0002
Unemployed to Unemployed	1.75	0.65	2.71	0.0068
Out of Labor Force to Unemployed	0.70	1.01	0.70	0.4870
Employed to Out of labor Force	3.73	0.81	4.59	0.0000
Unemployed to Out of Labor Force	1.89	1.10	1.71	0.0870
Out of Labor Force to Out of Labor Force	0.65	0.74	0.88	0.3774

The logistic regression is weighted because our data are characterized by choice-based sampling. The weightin scheme is described in Manski and Lerman (1977).

BEST COPY AVAILABLE

Table 3(d)
Estimated Coefficients from Weighted Logit*

Response Variable: D=1 if Control, 0 if ENP
Sample of Female Youths 170 controls and 346 ENPs

Variable	Coeff	Standard Error	T-Statistic	Prob> | T |
Intercept	-6.27	0.87	-7.19	0.0000
Fort Wayne, Indiana	0.71	0.53	1.35	0.1757
Jersey City, New Jersey	-0.03	0.52	-0.06	0.9487
Providence, Rhode Island	1.09	0.52	2.10	0.0357
Black	0.66	0.39	1.67	0.0958
Hispanic	0.65	0.45	1.44	0.1506
Race other than White, Black or Hispanic	-0.14	0.62	-0.22	0.8226
Age 19 to 21	-0.83	0.30	-2.73	0.0063
Less than Tenth Grade	-0.62	0.39	-1.58	0.1146
Tenth-Eleventh Grade	-0.68	0.34	-2.01	0.0045
One to Three Years College	0.27	0.45	0.60	0.548 1
Last Married 1-24 Months Prior to Random Assignment	2.04	0.82	2.48	0.0131
Last married More than Twenty-Four Months Prior to Random	0.29	1.30	0.22	0.8251
Assignment	0.61	0.47	1.32	0.1884
Single, Never married	-0.68	0.34	-2.01	0.0445
Children Less Than Six	1.71	0.59	2.90	0.0037
Unemployed to Employed	1.28	0.49	2.60	0.0094
Out of Labor Force to Employed	2.37	0.58	4.08	0.0000
Employed to Unemployed	3.09	0.68	4.55	0.0000
Unemployed to Unemployed	4.01	0.72	5.54	0.0000
Out of Labor Force to Unemployed	1.38	0.56	2.47	0.0134
Employed to Out of labor Force	0.37	1.02	0.36	0.7193
Unemployed to Out of Labor Force	1.14	0.54	2.11	0.0349
Out of Labor Force to Out of Labor Force	0.45	0.43	1.0s	0.2950
One Job Spell in Eighteen Months Prior to Random Assignment	1.11	0.45	2.50	0.0124
Two Job Spells	1.55	0.60	2.60	0.0094
Three or More Job Spells	2.59	0.51	508	0.0000
Receives Food Stamps but not AFDC	0.50	0.60	0.83	0.4088
Receives AFDC but not Food Stamps	1.42	0.40	3.51	0.0004
Receives both AFDC and Food Stamps				

The logistic regression is weighted because our data are characterized by choice-based sampling. The weighting scheme is described in Manski and Lerman (1977).

Table 4
Percent Correctly Classified Under the Regular Propensity Score Model

Demographic Group	Controls	ENPs	Within Sample Weighted	Population Weighted*
Percent Correctly Classified Under the Regular Propensity Score Model				
4dult Men	81.34	83.25	82.18	83.19
4dult Women	72.94	78.05	75.79	77.90
Male Youth	71.88	68.05	70.90	68.16
Female Youth	68.88	72.94	71.60	72.82
Percent Correctly Classified Under the SIPP Propensity Score Model				
Adult Men	80.97	85.25	83.04	85.12
Adult Women	83.89	85.20	84.60	85.16
Male Youth	78.75	78.65	78.73	78.65
Female Youth	75.43	81.72	77.2%	81.53
Percent Correctly Classified Under the No-Show Propensity Score Model				
Adult Men	67.51	62.05	65.16	65.16
Adult Women	66.71	63.72	65.36	65.36
Male Youth	69.06	66.92	68.34	68.34
Female Youth	67.80	61.01	65.49	65.49

The population weighted classification rate uses 0.03 as the weight for controls and 0.97 as the **weight** for ENI ; and SIPPs.

Figure 1 (a): Density of Estimated Regular Propensity Scores, Adult Men, Controls

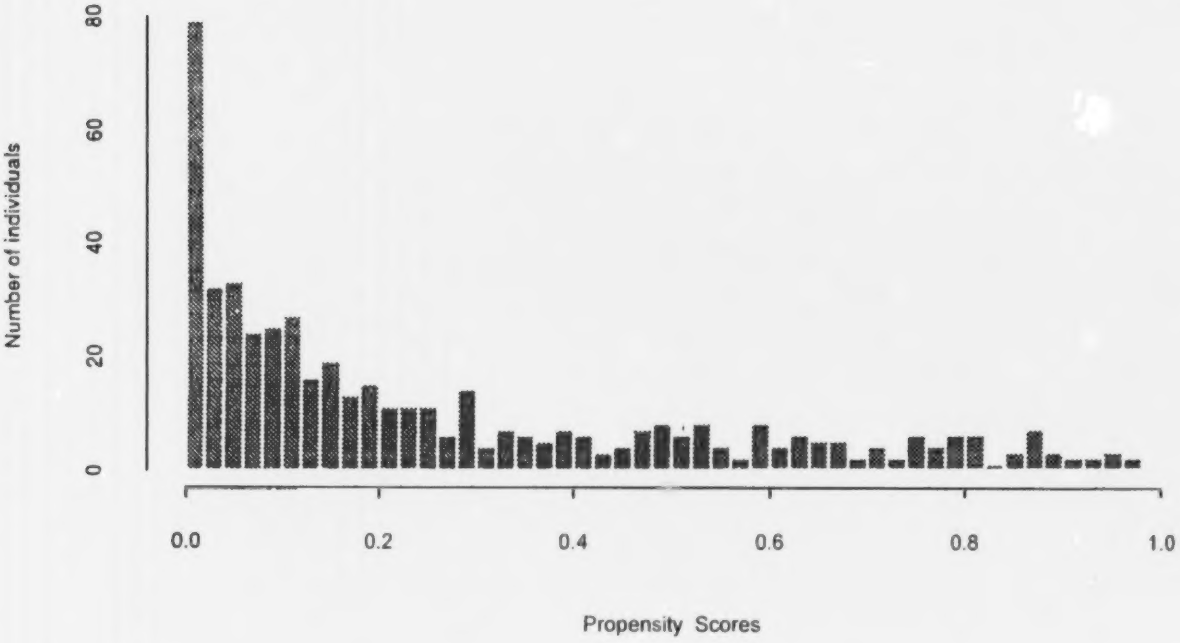


Figure I(a): Density of Estimate; Regular Propensity Scores, Adult Men, ENPs

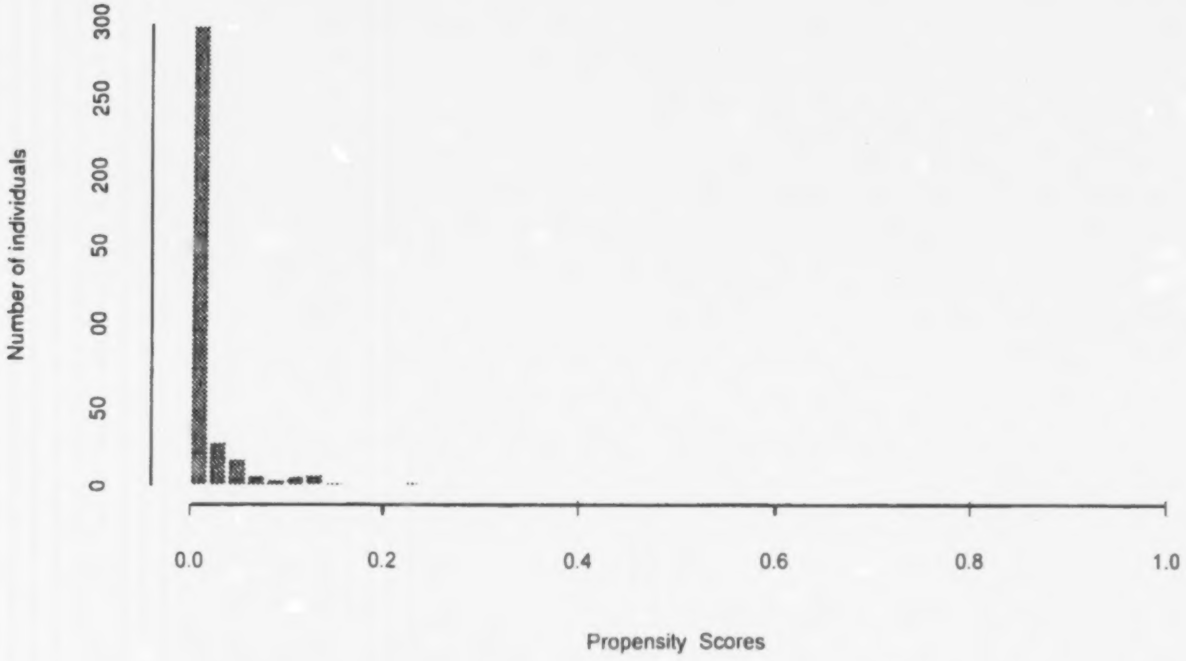


Figure 1 (b): Density of Estimated Regular *Propensity Scores, Adult Women, Controls*

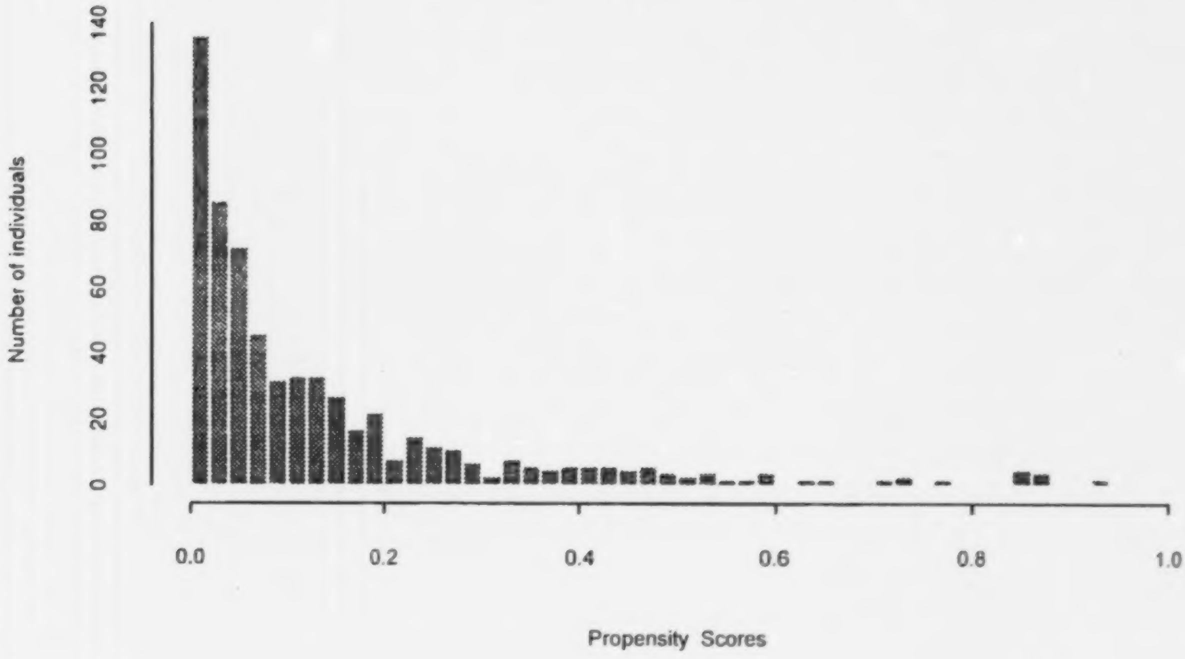


Figure I(b): Density of Estimated **Regular** Propensity Scores, Adult Women, ENPs

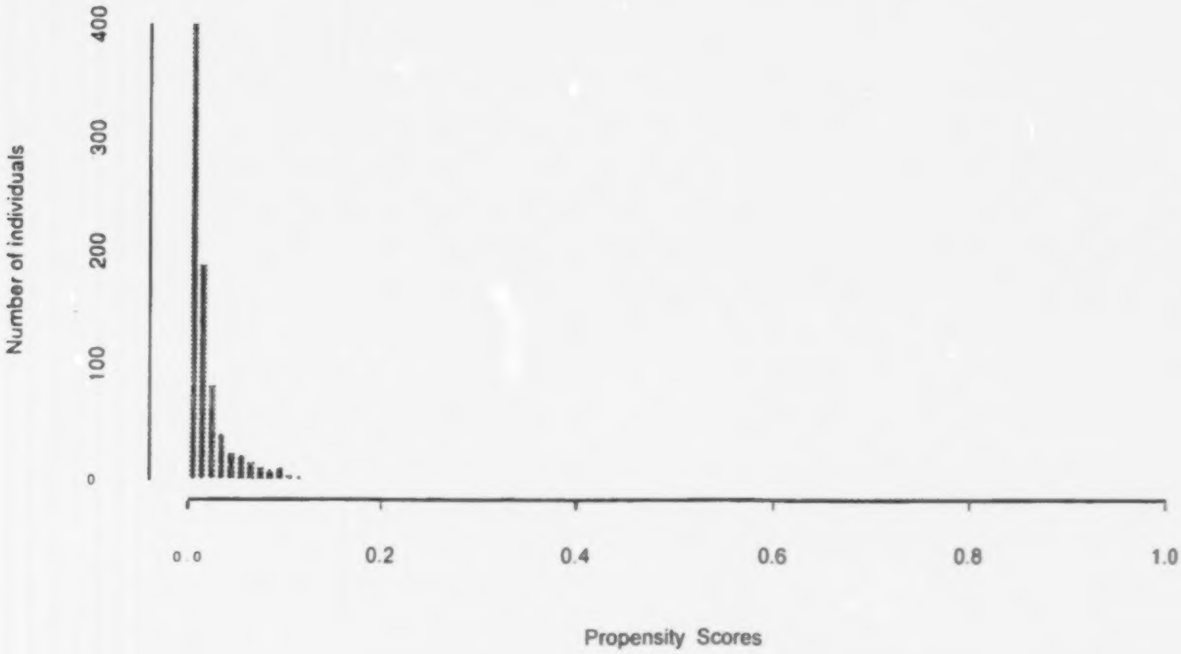


Figure 1 (c): Density of Estimated Regular Propensity Scores, Male Youth, Controls

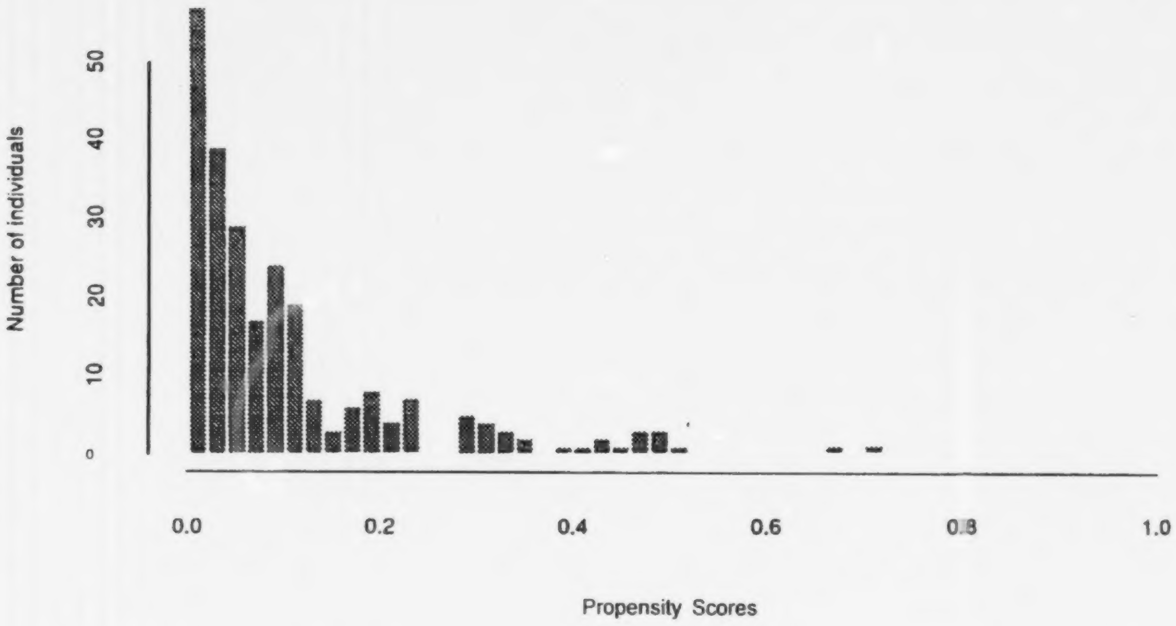


Figure 1 (c): Density of Estimated Regular Propensity Scores, Male Youth, ENPs

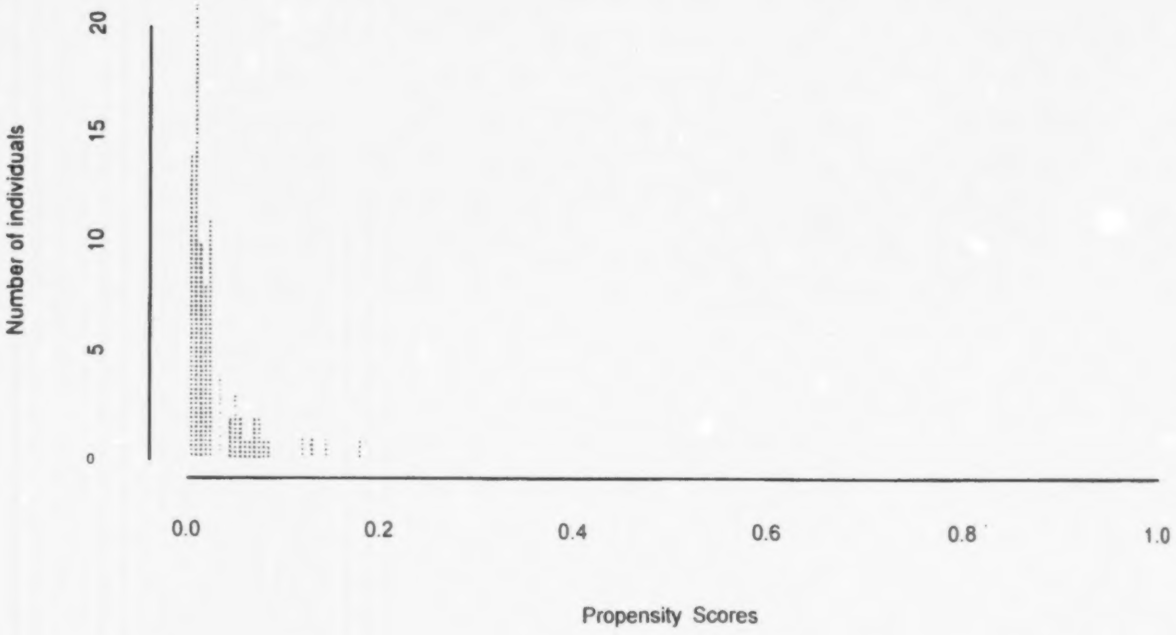


Figure 1 (d): Density of Estimated Regular Propensity Scores, Female Youth, Controls

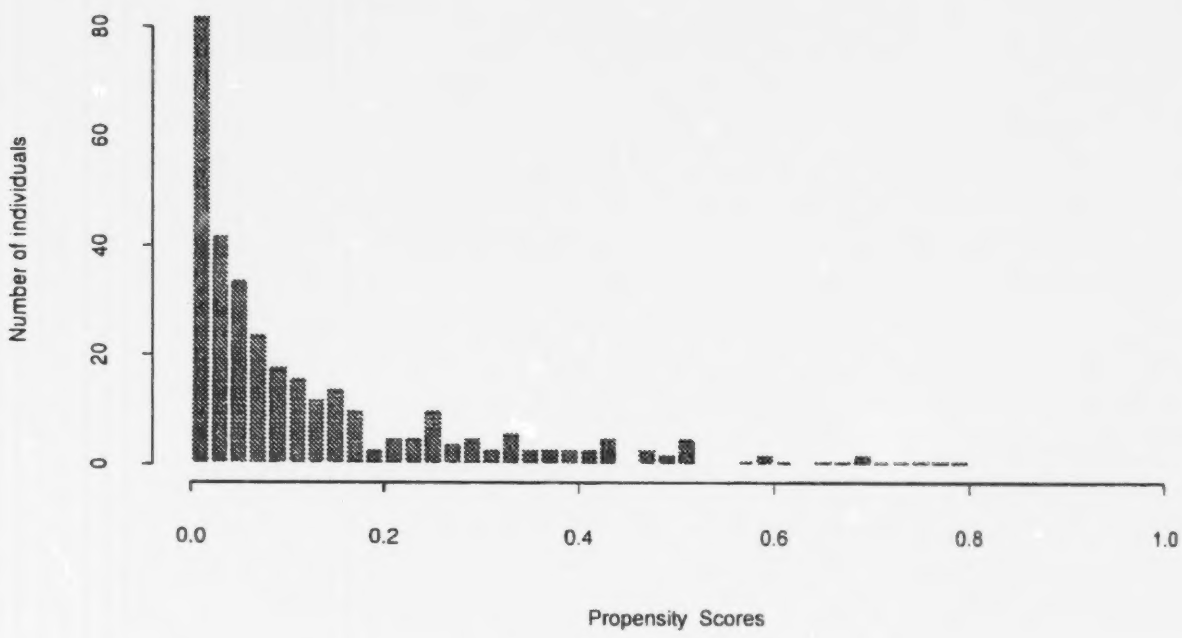
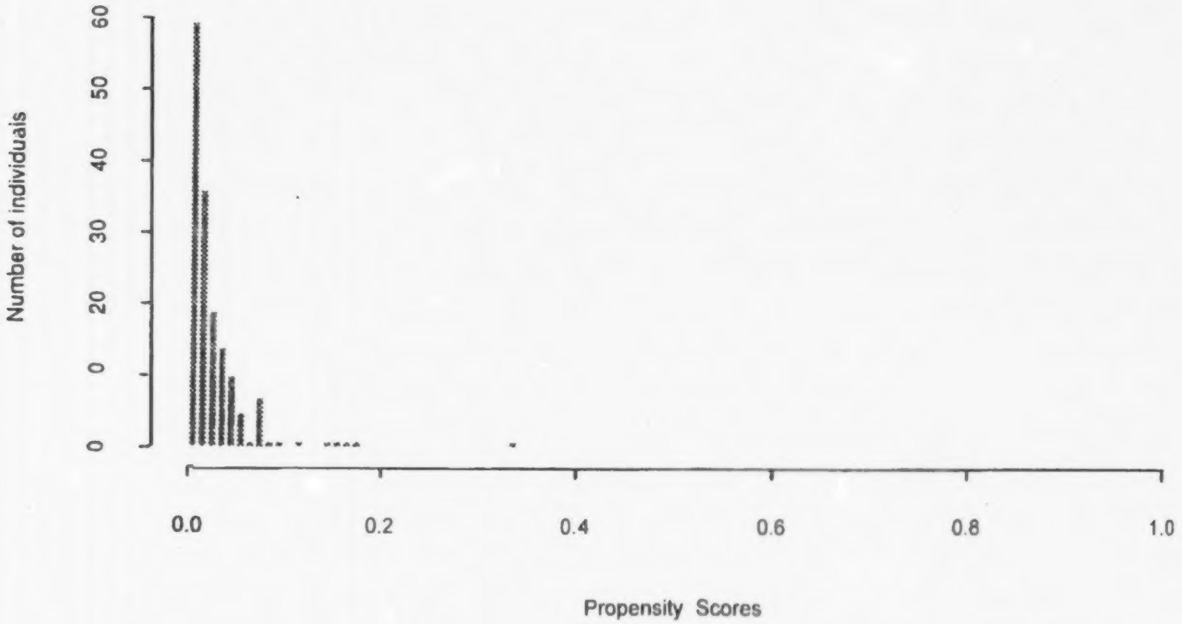


Figure 1 (d): Density of Estimated Regular Propensity Scores, Female Youth, ENPs



controls. In the empirical results reported below, we show that simple nearest neighbor estimates are usually more reliable when the range of P_i over which matching is performed is restricted to the overlapping support region. When nonparametric matching methods are used, the support problem has to be addressed because we require the density of P_i to be positive at each point where the matched earnings function is estimated. In an earlier joint work, we demonstrate that failure to match at comparable P values is a major source of what is commonly called selection bias.

3. Verifying the Conditions for Matching

As discussed in our companion paper, much of the statistical matching literature invokes the Rosenbaum and Rubin (1985) conditional independence assumptions. With data on a treatment group and on a nonexperimental comparison group, it is not possible to verify whether these assumptions are satisfied. Nevertheless, as described in Theorem 2 in Heckman, Ichimura, and Todd (1997a,b), consistent and asymptotically normal estimators estimate the parameter $E(Y_1-Y_0 \mid D=1)$ obtained even when the conditional independence assumption is violated but a weaker conditional mean independence assumption is met. We reject the conditional independence assumption but we do not reject mean independence for U_0, i.e.,

$$E(U_0 \mid D=1,P(Z)) = E(U_0 \mid D=0,P(Z)).$$

Heckman, Ichimura, Smith, and Todd (1997a) report favorable evidence on this hypothesis for results for adult men. We find that mean independence characterizes the X-adjusted residuals but not the raw data. This evidence demonstrates the importance of Theorem 2 in Heckman, Ichimura, and Todd in explaining our data. Mean independence is all that is required to implement kernel-smoothed matching. Conventional matching methods invoke stronger forms of conditional independence that are unnecessary and inconsistent with our data.

4. Evaluating the Performance of Different Matching Estimators

We consider the performance of some widely used statistical matching methods for parameter $E(Y_1 - Y_0 \mid D=1)$ which gives the mean impact of the program on groups of participants. We contrast the performance of several conventional matching methods in the literature to that of smoothed matching estimators.

In implementing the smoothed matching estimators, we use local linear regression smoothing instead of more conventional kernel smoothing. Local polynomial estimators have recently been shown by Fan (1993) to have superior properties to kernel estimators. They have a faster rate of convergence at boundary points and adapt better to different data design densities. The use of local linear regression methods in the evaluation problem is discussed in detail in Heckman, Ichimura, Smith, and Todd (1997). However, their use is not central to our analysis here, and kernel versions of each of our nonparametric estimators could also have been used.

We can define a rich class of matching estimators by

$$M = \sum_{i=1}^{N_1} [\, Y_{1i} - \sum_{j=1}^{N_0} W_{N_0,N_1}(i,j)\, Y_{0j}].$$

Matching estimators differ in the way they construct matches. Define a neighborhood $C(X_i)$ for person i in the participant sample. Neighbors in i are persons j for whom $X_j \in C(X_i)$. The persons matched to i are those people in set $\mathbf{A}_i$ for whom $\mathbf{A}_i = \{j \mid X_j \in C(X_i)\}$.

The nearest-neighbor matching estimator sets $W_{N_0 N_1}(i) = 1/N_1$ and picks for each i in sample $D = 1$

$$C(X_i) = \min_j \| X_i - X_j \|, j \in \{1, \ldots . N_0\}$$

where $\| \cdot \|$ is a norm. $\mathbf{A}_i$ is a singleton set except for ties that are broken by a random draw. The weighting scheme for the nearest-neighbor estimator is

$$\begin{array}{ll} W_{N_0,N_1} = 1 & \text{if } j \in A_i \\ \quad\quad\;\, = 0 & \text{otherwise.} \end{array}$$

Two versions of this method are (a) X_j may be reused for other matches (sampling with replacement) and (b) X_j may be discarded after it is matched (sampling without replacement). This method finds one person j for each i. The distance between person i and j can be substantial if $C(X_i)$ is not restricted.

A version of nearest-neighbor matching designed to avoid the problem of a substantial gap between i and j is "caliper" matching (Cochrane and **Rubin,** 1973). In this scheme, matches are made to person i only if

$$\| X_i - X_j \| < \varepsilon, \qquad j \in \{1, ..., N_0\}$$

where ϵ is a prespecified tolerance. Otherwise no match is undertaken, and person i is bypassed. In this scheme $C(X_i) = \{X_j \mid \| X_i - X_j \| < E\}$. If more than one person is in $\mathbf{A}_i$, then use the nearest neighbor under norm $\| \ \|$ to pick the matched person. Again, there may be sampling with or without replacement from the control sample. A variant of caliper matching selects one metric to caliper match and another to pick among elements in the control sample if more than one person qualifies. Nearest neighbors under the first metric are used if there is no observation that is within ε in the first metric (Rosenbaum and **Rubin,** 1985). This procedure guarantees that a match is made for each person in the control population. Again, like all of the methods, these can be applied with or without reuse of the observations in the comparison group.

The Mahalanobis metric is a common metric for the nearest-neighbor matching estimator. The metric used is

$$\|X_i - X_j\| = (X_i - X_j)' \Sigma^{-1} (X_i - X_j)$$

where $\sum$ is the covariance matrix in the pooled "within" covariance matrices formed for the samples D = 1 and D = 0 (i.e., covariances are formed removing within-sample means).

Kernel matching sets A, = $\{1,...N_0\}$ and

$$W_{N_0,N_1}(i,j) = \frac{K_{N_0}(X_j - X_i)}{\sum_{j=1}^{N_0} K_{N_0}(X_j - X_i)}$$

where kernel $K_{N_0}(s) = K(s/a_{N_0})$ andis a sequence of numbers that converges to zero. Kernel matching is a smooth method that reuses and weights all the comparison group observations for each person *i*.

The samples can be defined conditional on different strata. (Strata are discrete-valued conditioning variables.) The overall mean difference is

$$M = \frac{1}{N_1}\sum_{i=1}^{N_1}\left(Y_{1i} - \frac{1}{N_0}\sum_{j=1}^{N_0} W_{N_0,N_1}(i,j)\,Y_{0j}\right) = \frac{1}{N_1}\sum_{i=1}^{N_1}\left(Y_{1i} - \bar{Y}_{0i}\right)$$

for the strata where $\bar{Y}_{0i} = \frac{1}{N_0}\sum_{j=1}^{NO} W_{N_0,N_1}(i,j)Y_{0j}$.

The following matching procedures are evaluated in our empirical work:

1. Simple Propensity Score Nearest-Neighbor Matching: This match uses the n nearest neighbors in terms of the propensity score. $n = 1$ defines nearest-neighbor matching. We also consider simple average versions of nearest-neighbor matching, which average the outcomes of the nearest 5 and the nearest 10 neighbors to form the match.

2. Kernel-Smoothed Propensity Score Matching: This procedure forms a weighted average over the outcomes of individuals in the comparison group, using kernel regressions with observations with closer propensity scores receiving higher weight. Bandwidths equal to 0.02, 0.04, and the optimal plug-in bandwidth are used in estimation.

3. Mahalanobis Matching Within Propensity Score Calipers: this procedure restricts the set of possible matches to individuals with propensity scores within a specified range (called the caliper width). Within this range, the individual closest in terms of the Mahalanobis metric is chosen. If no individual is within the range, then the procedure matches with the person with the closest propensity score. We tried a fixed caliper width equal to $0.5\sqrt{(\sigma_{1p}^2 + \sigma_{0p}^2 /2}$, where σ_{1p}^2 and σ_{0p}^2 are the variance of the propensity scores within the participant and comparison group samples. We also tried variable caliper widths equal to the distance to the nth nearest propensity score, n = 5 and n = 10. Variable caliper widths guarantee a set of potential matches within the range.

4. Smoothed Mahalanobis Distance Matching Within Calipers: We narrow down the set of potential matches as described above using variable caliper widths of n = 5 and 10. We then construct a weighted average estimate of $E(Y_0 \mid D=1,P)$ using local linear regression smoothing. The bandwidth is equal to the distance to the nth nearest Mahalanobis metric, which insures that all the observations within the caliper width are used in the smoothing."

5. Regression-Adjusted Local Linear Matching: We remove the $g_0(X)$ effect from Y_1, and Y_0 and match on the adjusted outcomes as follows. Using data on persons who did not apply to the program, we estimate

$$Y_0 = g_0(X) + E(U_0 \mid P) + V$$

where $g_0(X)$ is assumed to be a linear function of X and $E(U_0 \mid P)$ is estimated nonparametrically by local linear regression. Consistent estimates of $g_0(X)$ are obtained because of the simultaneous estimation procedure. Adjusted outcomes for participants, $Y_1 - \hat{g}_0(T)$ are then matched with adjusted outcomes for nonparticipants, $Y_1 - \hat{g}_0(X)$. Local linear regression smoothing is used to estimate the matched outcome, $\hat{E}(Y_0 - \hat{g}_0(X) \mid D=1,P)$ with smoothing parameters equal to 0.02, 0.04, and the optimal bandwidth. This procedure is formally justified in **Heckman,** Ichimura, and Todd **(1997a,b).**

Table 5 reports estimates of the bias from a variety of matching estimators for the four demographic groups analyzed in this paper. Trimming at ten percent was performed to define the common support. (Trimming at a% uses support points for P where the estimated density of P is a% or higher). The first column in each table set reports the unadjusted mean difference between control earnings and participant earnings, quarter by quarter and averaged over the full thirty-six month

---

[10] We also tried fixed bandwidths equal to 0.02, 0.04, and 0.06. With the youth groups, we ran into problems using the fixed bandwidth, because it was often too small for some caliper groups. Very few observations were used in the computation of the match leading to large bias values.

period (t-6 to **t+6**) and the postprogram period (**t+l** to t+ 6). The second line from the bottom shows the bias as an estimate of the estimated postprogram impact from the program. The final row reports the bias as a percentage of the estimates of the bias from regression-adjusted local linear regression estimator given in the final column in each two table set. Thus, for adult men, the average postprogram bias in the raw means is **−$343** per month, which is more than five times the estimated program impact, and 1,805 percent of the local-linear-regression-adjusted bias of **−$23**.

For all demographic groups, the kernel-based smoothed estimators do as well, or better, than the other matching estimators and produce very small estimates of bias. Accounting for common support, the nearest-neighbor estimator does surprisingly well. (In **Heckman,** Ichimura, and Todd, **1995a,** we prove consistency of the estimator.) The Mahalonobis metric estimators that were widely used in the CETA evaluations (see Bryant and Rupp, 1986; and Dickinson et al., 1986) do amazingly poorly. Simple kernel-based matching on the *P* scores does well. The local linear regression version generally does even better. In general, the local linear regression estimator based on the optimal bandwidth does the best, and produces estimators with little bias, providing one adopts the average bias as the criterion of evaluation success.

5. The Importance of the Conditioning Variables

We now consider the effect of changing the conditioning set used in estimating the propensity scores. Previous nonexperimental evaluations of training program were based on much coarser data than our own. We seek to learn how the quality of the data influences the effectiveness of the matching estimator. It is not the case that the inclusion of additional regressor variables will always improve the effectiveness of a matching estimator. For example, introducing conditioning variables that perfectly classified applicants and nonapplicants would make matching impossible.

Our empirical results indicate that matching estimators perform best whenvariables describing recent unemployment histories, which were shown above to be significant determinants of participation in **Heckman** and Smith **(1994),** are included in the analysis. All of the estimates reported in Table 4 are based on participation models that use this information. Bias is substantially greater if only background demographic variables are used in estimating the propensity scores. Access to

Table 5(a)
Estimated Bias Under Alternative Nonparametric Matching Methods (Adult Men, Controls and ENPs)†

Quarter	Difference in Means	Nearest Neighbor Without Common support	Nearest Neighbor With Common support	Average Nearest 5 Neighbors	Average Nearest 10 Neighbors	Mahalanobis With Fixed Calipers	Mahalanobis With Nearest 5 Caliper Width	Mahalanobis With Nearest 10 Caliper Width
t+1	-425	61	-117	-119	-128	582	572	577
t+2	-353	-248	-126	-64	-65	688	678	683
t+3	-341	-204	-209	-75	-76	754	744	749
t+4	-287	66	-60	18	24	806	795	800
t+5	-309	121	5	65	81	816	807	811
t+6	-342	33	-11	49	59	779	769	773
Ave. t+1 to t+6	-343	-28	-86	-21	-18	738	728	732
as a % of:								
impact*	686%	56%	172%	42%	36%	1476%	1456%	1464%
LLR Adjusted	1491%	122%	374%	91%	78%	3209%	3165%	3183%
Quarter	Local Linear Propensity Score Matching (bw:0.02)	Local Linear Propensity Score Matching (bw:0.04)	Local Linear Propensity Score Matching (bw:optimal)	Smoothed Mahalanobis (5 nearest neighbor caliper and bandwidth)	Smoothed Mahalanobis (10 nearest neighbor caliper and bandwidth)	Regression Adjusted Local Linear Matching (bw:0.02)	Regression Adjusted Local Linear Matching (bw:0.04)	Regression Adjusted Local Linear Matching (bw:optimal)
t+1	-123	-143	-118	-246	-258	-61	-49	-93
t+2	-94	-98	-71	-185	-112	-102	-63	-16
t+3	-109	-98	-70	-286	-244	-165	-103	-61
t+4	-12	-7	23	-209	-118	-99	-43	-3
t+5	37	34	65	-194	-30	-54	-10	20
t+6	29	22	40	-193	-14	-42	-7	18
Ave. t+1 to t+6	-45	-28	-22	-219	-129	-87	-46	-23
as a % of:								
impact*	90%	96%	44%	438%	258%	174%	92%	46%
LLR Adjusted	196%	209%	96%	952%	561%	378%	200%	100%

† The variables in the Mahalanobis metric and in the regression-adjusted linear models are indicators for training site, race, education level, and marital status. In addition, there are indicator variables for age> [illegible] for the year of observation, for the yearly quarter of the observation, and for whether currently in training.

* The impacts for each demographic group are taken from Roselius (1995), Table 1, 18-month figures. The impacts in the table are comparable to t+1 to t+6.

Trimming at 10% is performed by first requiring that controls and ENPs have positive density at each point p (overlap support condition) and then trimming points below the 10% quantile.

BEST COPY AVAILABLE

Table 5(b)

Estimated Bias Under Alternative Nonparametric Matching Methods (Adult Women, Controls and ENPs)†

Quarter	Difference in Means	Nearest Neighbor Without Common Support	Nearest Neighbor With Common Support	Average Nearest 5 Neighbors	Average Nearest 10 Neighbors	Mahalanobis With Fixed Calipers	Mahalanobis With Nearest 5 Caliper Width	Mahalanobis With Nearest 10 Caliper Width
t+1	-28	66	4	10	0	290	276	280
t+2	18	84	50	43	27	361	353	357
t+3	31	87	31	30	22	382	369	373
t+4	50	71	25	32	19	399	385	389
t+5	50	49	26	31	17	412	398	402
t+6	30	102	5	12	-5	403	389	394
Ave. t+1 to t+6	25	76	24	26	13	376	362	366
as a % of:								
impact‡	66%	200%	63%	68%	34%	989%	953%	963%
LLR Adjusted	208%	633%	200%	217%	108%	3133%	3017%	3050%

Quarter	Local Linear Propensity Score Matching (bw:0.02)	Local Linear Propensity Score Matching (bw:0.04)	Local Linear Propensity Score Matching (bw:optimal)	Smoothed Mahalanobis (5 nearest neighbor caliper and bandwidth)	Smoothed Mahalanobis (10 nearest neighbor caliper and bandwidth)	Regression Adjusted Local Linear Matching (bw:0.02)	Regression Adjusted Local Linear Matching (bw:0.04)	Regression Adjusted Local Linear Matching (bw:optimal)
t+1	-4	-8	11	185	19	-25	-32	4
t+2	19	14	39	85	63	-7	-14	25
t+3	11	7	38	-4	40	-15	-21	20
t+4	5	1	27	66	3	-18	-24	18
t+5	5	1	20	-64	-31	-18	-24	14
t+6	-19	-23	3	-101	-105	-49	-56	-6
Ave. t+1 to t+6	3	-1	23	13	-2	-22	-28	12
as a % of:								
impact‡	8%	3%	61%	34%	5%	58%	14%	32%
LLR Adjusted	25%	8%	192%	108%	17%	183%	233%	100%

† The variables in the Mahalanobis metric and in the regression-adjusted linear models are indicators for training site, race, education level, and marital status. In addition, there are indicator variables for age > 25, for the year of observation, for the yearly quarter of the observation, and for whether currently in training.

‡ The impacts for each demographic group are taken from Roselius (1995), Table 1, 15-month figures. The impacts in the table are comparable to t+1 to t+6.

Trimming at 10% is performed by first requiring that controls and ENPs have positive density at each point p (overlap support condition) and then trimming points below the 10% quantile.

BEST COPY AVAILABLE

Table 5(c)

Estimated Bias Under A ernative
Nonparametric Matching Methods
(Male Youth, Controls an ENPs)†

Quarter	Difference in Means	Nearest Neighbor Without Common Support	Nearest Neighbor With Common Support	Average Nearest 5 Neighbors	Average Nearest 10 Neighbors	Mahalanobis With Fixed Calipers	Mahalanobis With Nearest 5 Caliper Width	Mahalanobis With Nearest 10 Caliper Width
t+1	-80	131	38	-23	-36	388	384	386
t+2	-26	246	170	4	12	432	429	431
t+3	-40	181	77	-39	-24	506	503	505
t+4	-11	267	151	20	12	584	580	583
t+5	46	187	148	91	91	615	611	613
t+6	24	-69	37	50	72	634	631	633
Ave. t+1 to t+6	-14	157	104	17	21	526	523	525
as a % of:								
impact*	36%	403%	261%	44%	54%	1349%	1341%	1346%
LLR Adjusted	34%	383%	254%	41%	51%	1283%	1276%	1280%

Quarter	Local Linear Propensity Score Matching (bw:0.02)	Local Linear Propensity Score Matching (bw:0.04)	Local Linear Propensity Score Matching (bw:optimal)	Smoothed Mahalanobis (5 nearest neighbor caliper and bandwidth)	Smoothed Mahalanobis (10 nearest neighbor caliper and bandwidth)	Regression Adjusted Local Linear Matching (bw:0.02)	Regression Adjusted Local Linear Matching (bw:0.04)	Regression Adjusted Local Linear Matching (bw:optimal)
t+1	8	6	-4	83	8	6	5	5
t+2	48	49	48	345	66	25	28	36
t+3	20	26	17	269	9	-19	-12	-14
t+4	60	64	58	15	-75	31	36	49
t+5	106	111	123	195	94	80	84	114
t+6	67	76	85	165	125	25	32	55
Ave. t+1 to t+6	52	55	54	179	38	25	29	41
as a % of:								
impact*	133%	141%	138%	459%	97%	64%	74%	105%
LLR Adjusted	127%	134%	132%	437%	93%	61%	71%	100%

†The variables in the Mahalan is metric and in the regression-adjusted linear models are dicators for training site, race, education level, and marital rus. In addition, there are indicator variables for age > 5 for the year of observatic for the yearly quarter of the observation, and for whether currently in training.

‡ The impacts for each demographic group are taken from Roselius (1995), Table 1, 18-month figures. The impacts in the table are comparable to t+1 to t+6.

Trimming at 10% is performed by first requiring that controls and ENPs have positive density at each point p (overlap support condition) and then trimming points below the 10% quantile

BEST COPY AVAILABLE

Table S(d)

Estimated Bias Under Alternative Nonparametric Matching Methods (Female Youth, Controls and ENPs)†

Quarter	Difference in Means	Nearest Neighbor Without Common Support	Nearest Neighbor With Common Support	Average Nearest 5 Neighbors	Average Nearest 10 Neighbors	Mahalanobis With Fixed Calipers	Mahalanobis With Nearest 5 Caliper Width	Mahalanobis With Nearest 10 Caliper Width
t+1	7	59	-32	27	20	218	208	213
t+2	35	89	17	66	58	284	273	277
t+3	63	141	77	100	90	351	340	345
t+4	7	69	-4	12	3	353	343	347
t+5	43	97	23	34	36	331	321	325
t+6	12	78	26	7	18	349	338	342
Ave. t+1 to t+6	28	89	18	41	38	314	304	308
as a % of:								
impact‡	467%	1483%	300%	683%	633%	5233%	5067%	5133%
LLR Adjusted	97%	307%	62%	141%	131%	1083%	1048%	1062%

Quarter	Local Linear Propensity Score Matching (bw:0.02)	Local Linear Propensity Score Matching (bw:0.04)	Local Linear Propensity Score Matching (bw:optimal)	Smoothed Mahalanobis (5 nearest neighbor caliper and bandwidth)	Smoothed Mahalanobis (10 nearest neighbor caliper and bandwidth)	Regression Adjusted Local Linear Matching (bw:0.02)	Regression Adjusted Local Linear Matching (bw:0.04)	Regression Adjusted Local Linear Matching (bw:optimal)
t+1	24	26	37	-68	-74	6	8	18
t+2	61	66	79	-177	-182	42	46	58
t+3	97	102	113	-213	-189	55	59	70
t+4	21	28	29	-193	-189	-21	-14	-8
t+5	47	55	65	-53	-157	7	15	27
t+6	20	25	38	-203	-216	-10	-6	8
Ave. t+1 to t+6	45	50	60	-151	-168	13	18	29
as a % of:								
impact‡	750%	833%	1000%	2517%	2800%	217%	300%	483%
LLR Adjusted	155%	172%	207%	521%	579%	45%	62%	100%

The variables in the Mahalanobis metric and in the regression-adjusted linear models are indicators for training site, race, education level, and marital status. In addition, there are indicator variables for age>25 for the year of observation, for the yearly quarter of the observation and for whether currently in training.

‡ The impacts for each demographic group are taken from Roselius (1995), Table 1, 18-month figures. The impacts in the table are comparable to t+1 to t+6.

Trimming at 10% is performed by first requiring that controls and ENPs have positive density at each point p (overlap support condition) and then trimming points below the 10% quantile.

BEST COPY AVAILABLE

information on earnings in the previous year greatly improves the performance of the matching estimator.

We estimate propensity scores using four different sets of variables that differ in the type of data used to describe recent labor force dynamics. The names ascribed to the different conditioning sets are coarse I, coarse II, coarse III, and coarse IV. The regressors included in each model and their relationship to the regular propensity score model analyzed previously are summarized below.

> **Base Regressor Set:** Site, age, race, education, marital status, presence of children under 6 years old.
>
> **Regular Propensity Scores:**
>
> > Adult men-In addition to the base regressors, include indicators for whether recently in vocational training, for the number of job spells in 18 months prior to random assignment, for the most recent labor force transition pattern, and a variable indicating the number of household members.
> >
> > Adult women-In addition to the base regressors, include indicators for number of recent labor force transitions, for the most recent labor force transition pattern, and for recent vocational or classroom training.
> >
> > Male youth-In addition to base regressors, include indicators for the most recent labor force transition pattern.
> >
> > Female youth-In addition to base regressors, include indicators for the most recent labor force transition pattern, for the number of recent job spells, and for whether receiving food stamps or AFDC.
>
> **Coarse I Propensity Scores:** In addition to the base regressors, contain previous annual earnings.
>
> **Coarse II Propensity Scores:** In addition to the base regressors, contain previous annual earnings and the labor force transition from month t-12 to t-6.
>
> **Coarse III Propensity Scores:** In addition to the base regressors, contain previous annual earnings and labor force status at the time of randomization/eligibility determination.
>
> **Coarse IV Propensity Scores:** Contain only the base regressor variables.

Table 6
Bias From Local Linear Regression Matching Estimator

(a) Adult Men Using Optimal Band Width								
Quarter	Regular	Coarse 1	Coarse 2	Coarse 3	Coarse 4	Site Mismatch	SIPP	No-Show
t+1	-93	-62	-227	126	-366	-234	103	70
t+2	-16	21	-12.5	145	-269	-185	193	34
t+3	-61	40	-86	138	-257	-174	225	23
t+4	-3	53	-5	142	-184	-146	135	5
t+5	20	11	-3	111	-194	-171	*	4
t+6	18	-19	-43	85	-232	-148	*	-12
Avg. Post-Program	-23	7	-82	124	-250	-176	164	21

(b) Adult Women Using Optimal Band Width								
Quarter	Regular	Coarse 1	Coarse 2	Coarse 3	Coarse 4	Site Mismatch	SIPP	No-Show
t+1	4	-30	-56	13	-117	-120	48	59
t+2	25	4	-1	46	-62	-73	122	40
t+3	20	7	17	48	-38	-77	99	23
t+4	18	51	13	85	-21	-71	98	42
t+5	14	62	4	97	-11	-54	*	35
t+6	-6	38	-11	69	-35	-56	*	36
Avg. Post-Program	12	24	-6	60	-47	-75	92	39

The regular propensity score model includes indicator variables for site, age, race, education, marital status, the presence of children less than 6, most recent labo r force transition, number of job spells in the last 18 months, whether in vocational training at the ti of enrollment or ever ad vocational training, and a variable indicating the number of household members.

Coarse 1 propensity score model includes indicator variables for site, age race, education, marital status, the presence of children less than 6 and previous annual earnings.

Coarse 2 propensity score model includes indicator variables for site, age, race, education, marital status, the presence of children less than 6, previous annual earnings and labor force transition from months t-12 to t-5

Coarse 3 propensity score model includes indicator variables for site, age, race, education, marital status, the presence of children less than 6, previous annual earning and current labor force status

Coarse 4 propensity score model includes indicator variables for site, age, race, education marital status, and the presence of children less than 6

Site mismatch propensity score model is the same as the regular propensity score model

SIPP propensity score model includes indicator variables for age, race, education marital status, the presence of children less than 6 and most recent labor force status.

No-Show propensity score model includes indicator variables for site, age, race, education, marital status, the presence of children less than 6 age>25 for the year of observation, and the yearly quarter of the observation.

Trimming at 10% is performed by first requiring that both groups (either controls and ENPs or controls and SIPPs or controls and no-shows) have positive density at each point p (overlap support condition) and then trimming points below the 10% quantile.

* Data not available to compute for this period

Regular: Optimal propensity score used based on Heckman and Smith (1994)

SIPP: SIPP JTPA eligibles matched with JTPA controls (local linear regression with optimal bandwidth used to compute the estimates)

Site Mismatch: Controls from Providence and Jersey City matched with ENPs from Corpus Christie and Fort Wayne (local linear regression with optimal bandwidth used to compute the estimates)

No-Show: Persons who enrolled in JTPA but who dropped out before receiving services (local linear regression with optimal bandwidth used to compute the estimates) based on Table 7

BEST COPY AVAILABLE

Table 6(2)

Bias From Local Linear Regression Matching Estimator

(a) Male Youth Using Optimal Band Width

Quarter	Regular	Coarse 1	Coarse 2	Coarse 3	Coarse 4	SIPP	No-Show
t+1	5	50	-117	74	-96	-79	108
t+2	36	121	-28	131	-36	-14	43
t+3	-14	81	-96	101	-78	30	91
t+4	49	162	-104	170	-99	103	123
t+5	114	219	-9	235	-29	•	91
t+6	55	146	-49	181	-38	•	88
Avg. Post-Program	41	130	-67	149	-63	10	91

(b) Female Youth Using Optimal Band Width

Quarter	Regular	Coarse 1	Coarse 2	Coarse 3	Coarse 4	SIPP	No-Show
t+1	18	62	-7	95	-37	58	51
t+2	58	83	48	113	5	65	38
t+3	70	93	48	119	17	132	69
t+4	-8	51	-17	55	-26	116	25
t+5	27	82	20	85	22	•	36
t+6	8	80	-24	70	-3	•	74
Avg. Post-Program	29	75	11	90	-4	93	49

The regular propensity score model includes indicator variables for site, age, race, education, marital status, the presence of children less than 6, most recent labor force transition, number of job spells in the last 18 months, whet [illegible] in vocational training at the time of enrollment or ever had vocational training and a variable indicating the number of household members.

• Coarse 1 propensity score model includes indicator variables for site, age, race, education. marital status, the presence of children less than 6 and previous annual earnings.

Coarse 2 propensity score model includes indicator variables for site, age, race, education, marital status. the presence of children less than 6. previous annual earnings and labor force transition from months t-12 to t-6

Coarse 3 propensity score model includes indicator variables for site, age, race, education, marital status, the presence of children less than 6, previous annual earnings and current labor force status.

Coarse 4 propensity score model includes indicator variables for site, age, race, education, marital status, and the presence of children less than 6.

SIPP propensity score model includes indicator variables for age, race, education, marital status, the presence of children less than 6 and most recent labor force transition.

• No-Show propensity score model includes indicator variables for site, age, race, education, marital status, the presence of children less than 6, age>25 for the year of observation, and the yearly quarter of the observation

Trimming at 10% is performed by first requiring that both groups (either controls and ENPs or controls and SIPPs or controls and no-shows) have positive density at each point p (overlap support condition) and then trimming points below the 10% quantile.

• Data not available to compute for this period

•• Regular: Optimal propensity score used based on Heckman and Smith (1594)

•• SIPP: SIPP JTPA eligibles matched with JTPA controls (local linear regression with optimal bandwidth used to compute the estimates)

•• No-Show: Persons who enrolled in JTPA but who dropped out before receiving services (local linear regression with optimal bandwidth used to compute the estimates) based on Table 7.

BEST COPY AVAILABLE

Table 6 shows the importance of building a good model of program participation and demonstrate which variables are important and which are not in making good matching variables. Results are reported for the local linear regression estimator ("Regular") and for the four coarse matching variables. (The "SIPP" results are discussed in the next section.)

For adult males, the choice of the matching variables matters greatly. Failure to control for earnings or employment histories (Coarse 4) produces a badly biased estimator. The same cannot be said for the other demographicgroups. Consistent with the evidence reported in Heckman and Smith (1994), controlling for previous annual earnings on top of the base set of variables does a surprisingly good job for adult men. This is also true for all of the other groups.

The safest generalization from these tables is that kernel-based propensity score matching works well if it is based on a good model of program participation. For adult men, a good model requires either earnings or unemployment histories. For adult females or female youth, a variety of models are effective, and there is little evidence of selection bias once one controls for basic demographic characteristics and once one matches persons on comparable X values.

6. Using the Survey of Income and Program Participation as a Comparison Group

The SIPP data are richer than those used for comparison groups in the past. Several features of the SIPP make it an attractive source for constructing comparison groups.” It is longitudinal, has monthly observations on earnings and employment, the samples are large, and it contains enough information to determine eligibility status for JTPA, which is difficult to determine (see Devine and Heckman, 1994). We use the 1988 SIPP panel data set with observations that span from October 1987 to December 1989. For any individual, the length of the panel is twenty-four months. The 1988 panel covers the time period of the JTPA National Experiment at the four long baseline sites where random assignment took place from November 1987 to September 1989. The data contain

---

[11] In many of the studies reviewed in Barnow's (1987) survey of the CETA nonexperimental studies, program impacts were estimated using matched individuals from the Current Population Survey (CPS). Problems encountered in using this source of data included the inability to determine eligibility status for the program under consideration, time alignment differences between the participants and the comparison groups, and imprecise labor force histories. See Heckman (1995).

earnings and labor force histories and information on participation in JTPA, although few JTPA participants are found in the data in any month."

A drawback in using SIPP as a comparison group for the JTPA data is that it is a broad, nationally representative sample, whereas the JTPA data are from predominantly small to medium cities. The difference in geographic coverage induces differences in mean earnings levels that cannot be explained solely by differences in employment rates. Mean earnings are substantially different between the JTPA control group and SIPPs even after roughly matching on regions of residence or local unemployment rates in regions of residence. However, employment rates and labor force participation rates are similar between the two groups, suggesting that the observed earnings differences are due to cost-of-living differences or to differences in survey questions on earnings.

The final columns of Table 6 demonstrate the importance of matching people in the same labor market using the same questionnaire. LaLonde's (1986) influential study used comparison groups that are mismatched geographically and in terms of the questionnaires administered to participants and controls. Except for male youth, the estimated biases in using the SIPP data in place of ENP data are large. A substantial part of what is often interpreted as selection bias arising from individual participation decisions arises from geographical mismatch of questionnaires. Roselius (1995) reports estimates from a number of different SIPP samples that confirm this finding. Even closely regionally aligned SIPP samples produce estimates that are substantially biased. Smith (1995) investigates the separate roles of survey instruments and geographical mismatch in accounting for ENP-SIPP discrepancies. His lessons are of general interest because the JTPA data were collected in the format of the widely used NLSY data. About two-thirds of the discrepancy between the two data sources that produces the bias reported in Tables 6(a)-(d) is due to discrepancies in the questionnaires. The rest is due to geographical mismatch of participants and control group members.

---

[12] JTPA is called by different names in different localities. Persons participating in the JTPA program may not recognize it by that name, thus producing an understatement of true participation rates.

7. Summary of Section IV

This section of the paper uses experimental data from the JTPA program to evaluate the performance of a variety of matching estimators, including the kernel-smoothed methods justified in our theoretical analysis. It draws substantially from papers with Ichimura, smith and Todd. We find that an evaluation method based on (1) constructing a model that best predicts program participation and (2) using regression-adjusted kernel-smoothed matching on the propensity score obtained from the first stage, produces an estimator that is close to an estimator that would be produced by an ideal random experiment. Alignment of persons to a common support of the X variables is essential to the successful performance of matching as an evaluation method.

We also find that a substantial portion of what is conventionally regarded as selection bias in the nonexperimental program evaluation literature arises from a mismatch of participants and comparison groups in terms of geography and questionnaires. Our enthusiasm for matching is somewhat dampened when we look at the pointwise bias rather than the bias estimated over some interval. This bias is substantial (see **Heckman,** Ichimura amd Todd, 1996). Method matters but data matter, too, and both are equally important.

Appendix A

A More General Nonparametric Model for Conditional Means

Consider a more general model than the one used in section II of the text. Now the outcome equations are nonparametric. $\mu_1(X)$ is $E(Y_1 \mid X)$. $\mu_0(X)$ is $E(Y_0 \mid X)$.

(A-la) $$Y_0 = \mu_0(X) + U_0$$

(A-lb) $$Y_1 = \mu_1(X) + U_1$$

where $E(U_0 \mid X) = 0$ and $E(U_1 \mid X) = 0$. In the traditional regression setting, $\mu_0(X) = X\beta_0$ and $\mu_1(X) = X\beta_1$. Observed outcome Y can be written as

$$Y = DY_1 + (1-D)Y_0$$

i.e., it is either Y_1 or Y_0. If we insert (A-la) and (A-lb) into this expression, we obtain

(A-2) $$Y = \mu_0(X) + D(\mu_1(X) - \mu_0(X) + U_1 - U_0) + U_0.$$

This is a "two regime" or "switching regression" model (see Quandt, 1972). Labor economists call it a Roy model (see, e.g., Willis and Rosen, 1979, and **Heckman** and **Honoré**, 1990).

The term multiplying D is the gain. The gain has two components: $\mu_1(X) - \mu_0(X)$ — the gain for the average person-and $U_1 - U_0$— the idiosyncratic gain for a person. In this notation, then, the average gain is

$$E(\Delta \mid X) = \mu_1(X) - \mu_0(X).$$

The effect of treatment on the treated is

$$\mathrm{E}(\mathrm{A} \mid X,D=1) = \mu_1(X) - \mu_0(X) + E(U_1 - U_0 \mid X,D=1) .$$

The latter expression differs from the former by the additional term $E(U_1 - U_0 \mid X,D=1)$. This tells you how much the *gain of participants* differs from the average gain that would be experienced by the entire population. This is the gain to the movers from going from "0" to "1." Even though on average persons gain $\mu_1(X) - \mu_0(X)$, for *participants* the gain will be different. We may rewrite equation (A-2) in terms of these parameters:

(A-3) $$Y = p_{,,}(X) + D[E(\Delta \mid X)] + \{ U_0 + D(U_1 - U_0)\}$$

and

(A-4) $$Y = \mu_0(X) + D[E(A \mid X,D=1)] + \{ U_0 + D(U_1 - U_0 - E(U_1 - U_0 \mid X,D=1))\}.$$

This notation is dense. It can be simplified back to the notation used in Section II of this paper. Let $\mu_0(X)$ = a(X), $\beta(X) = \mu_1(X) - \mu_0(X) + U_1 - U_0$. Let the mean of β conditional on X be $E(\beta(X) \mid X)$. Let $\epsilon = U_1 - U_0$. Let $U_0 = U$. For notational simplicity, the dependence of α and β on X is left implicit: $\alpha(X) = \alpha$, $\beta(X) = \beta$. Then (A-4) can be written as a dummy-variable regression, like we encountered in the text:

$$Y = \alpha + D\beta + U.$$

Now, however, even after conditioning on X, β varies in the population due to individual heterogeneity in response to treatment. Equation (A-3) can be written as

$$Y = \alpha + D\bar{\beta} + \{U + D\epsilon\}. \tag{A-3}$$

$\bar{\beta}$ is the effect of placing an average person in the population at large into the program. (In general, $\bar{\beta}$ is a function of X, $\bar{\beta}(X)$. Again, the conditioning is left implicit). The effect of treatment on the treated is

$$E(A \mid X, D=1) = E(\beta \mid X, D=1) = \bar{\beta} + E(\epsilon \mid X, D=1) = \beta^*.$$

Again, I leave implicit the dependence of β^* on X. X plays no important role here because it is assumed to be mean independent of U, $E(U \mid X) = 0$. The analogue to equation (A-4) is

$$Y = \alpha + D\beta^* + \{U + D(\epsilon - E(\epsilon \mid D=1))\}.$$

All of the analysis in the text applies to this more general framework.

Appendix B

Comparison With Conventional Random Coefficient Models

The random coefficients model captures the idea that the response of a person to a variable may differ across persons. (See Judge et al., 1980, for a reference to such models.) Thus, in a traditional linear regression model

$$Y_i = X_i \underline{\gamma}_i + U_i$$

we explicitly allow the coefficient to bear a subscript i to identify variables that differ among persons. This is an intuitively attractive model because it allows for person-specific responses to changes in X.

The mean value of γ_i , is $\bar{\gamma} = E(\gamma_i)$. We assume that this mean is finite. Assuming that X_i, γ_i and U_i are statistically independent of each other, and $E(U_i) = 0$, we may write $Y_i = X_i\bar{\gamma} + \{U_i + X_i\epsilon_i\}$ where $\epsilon_i = \gamma_i - \bar{\gamma}$ which is in the form of a common-coefficient model.

The composite error term has mean zero, and conditional on X_i it has a variance component added to the usual error term. Thus

$$\mathrm{Var}\,(U_i + X_i\epsilon_i \mid X_i) = Var(U_i\ ) + Var(\epsilon_i)X_i^2.$$

In this conventional model, X_i could be a dummy variable for "treatment," in which case we have a "components of variance" model. Persons receiving treatment have additional variability in outcomes due to their variability in the response to treatment.

The model in the text for the case of $\epsilon = U_1 - U_0 \neq 0$ has D mean independent of $\epsilon(E(\epsilon \mid X,D=1) = 0)$. But unlike the traditional random coefficient model, there is no presumption the error component that ϵD is (mean) independent of U, or that D is (mean) independent of U. Thus, there may be a selection problem $E(U \mid X,D=1) \neq 0$. However, if $E(\epsilon \mid X,D=1) = 0$ the only econometric problem in estimating the parameters of (5') or (6') arises from mean dependence between U and D. the first component of the error term in (5') or (6').

Evaluating the School-to-Work Opportunities Act of 1994

Robert A. Moffitt
Department of Economics
The Johns Hopkins University

I. Introduction and Background

The School-to-Work Opportunities Act of 1994 (STWOA) is a major piece of legislation in education reform. The Act provides seed money to States and their local partners-business, labor, government, education, and community organizations-to develop school-to-work systems. The Act does not necessarily intend to encourage the development of new programs per se, although that will probably occur, but at a minimum, to encourage States to improve the coherence and organization of their existing systems in ways that will result in overall improvement.

Evaluation of the Act presents major technical and conceptual challenges. A plethora of types of school-to-work programs have been in operation around the country in different States and localities prior to the Act and will continue to exist after the Act is implemented. The tremendous diversity of program types implies that, in evaluation terminology, there will be an enormous number of "treatments" to evaluate. Measuring the net impacts of all of them, or even a significant number, will be difficult.

Because the Act is intended only as an incremental improvement to an existing school-to-work system, an evaluation should aim to determine the effects of the Act relative to the existing system; that is, to determine the incremental change in outcomes that can be ascribed to the Act. This is a more achievable and realistic goal than attempting to evaluate the effects of school-to-work programs relative to no program or relative to no system. In addition, while an obvious major goal should be to determine whether specific program changes implemented because of the Act do or do not improve student outcomes, and by how much, the larger goal should be to provide information on whether the States successfully use the funds to improve statewide coherence and to implement specific program improvements-that is. implementation success should be a goal of the evaluation. The level of success in this dimension should also be measured relative to the dollar inputs provided.

Given the diversity of the existing system and of the system that will emerge from the Act, a critical preliminary step in any evaluation should be a formal characterization of types of programs (treatments). A program typology should be developed identifying the key characteristics along which programs differ. This components-based typology should be capable of capturing most or all of the programs in operation around the country. The typology should also include a dimension for intensity, or extensiveness, of the treatment, including the size of the population served relative to the eligible population and possibly the funding level.

A major design decision concerns the choice of a random assignment or observational approach to estimating net impact, or a combination of the two. The strengths of a random assignment evaluation in general are: (1) improvement in the credibility of the results by reducing the threats to internal validity and (2) relatively low cost and relatively few data needs. The weaknesses of a random assignment evaluation are: (1) difficulties of implementation in the field if there is resistance to randomization, (2) ethical or legal barriers to denial of treatment, (3) difficulties in obtaining estimates for more than a small number of treatment types relative to the variety being implemented across the country, (4) a danger of possibly little external validity or generalizability, (5). difficulties in designing such evaluations in a way that yields information on scale effects, entry effects, and program participation effects, and (6) a risk of a noninformative evaluation.

The strengths of a proper observational study are: (1) information on a wide variety of treatments under a wide variety of circumstances can be obtained, (2) ethical problems are reduced and implementation problems may be reduced, (3) external validity and generalizability are relatively strong, and (4) information on scale effects, entry effects, and program participation effects can be obtained with a proper design. The weaknesses of observational analysis are: (1) a risk of lack of internal validity (selection bias) and (2) relatively high-cost data collection requirements. The data-collection needs are greater in part because more preprogram information and information on student and school characteristics must generally be collected for a credible evaluation, and also because information on why different programs are offered in different locations and on the numbers and characteristics of both participants and nonparticipants must be collected.

Given the relative advantages and disadvantages of experimental and observational analysis, a balanced strategy could be devised by conducting a series of randomized trials in a selected number of areas and for a selected number of programs (provided denial of services is possible), and by simultaneously collecting the minimum dataset (MDS) on a sufficient number of areas and programs to conduct a proper observational analysis on a large fraction of the population. A successful integrated evaluation design would make use of the advantages of both types of analyses and, through a synthesis of the results, yield a comprehensive evaluation of the Act.

Therefore, the following four steps are recommended for an evaluation of the Act:

- ***Develop a components-based program typology.*** Develop a classification system that minimizes black box categories and maximizes the use of attributes, characteristics, and components that can be compared across programs.

- ***Develop a data collection strategy for learning what types of programs are being implemented and what their attributes are.*** Part of this goal can be achieved using the data from the reporting requirements of the Act. Additional information can be obtained from the indepth study of sixty-four schools in eight states that is already under way. However, an additional sample of schools should be considered in order to obtain a representative picture of the different types of programs that are being implemented across the country. The results of this information-gathering effort can also be used to modify the program typology.

- ***Develop an immediate plan for a major data collection effort on schools and students to provide the basis for a comprehensive observational analysis.*** As soon as feasible, a sample of students and schools, designed either to be representative or to be stratified by components of the program typology, should be drawn with the largest possible number of schools and students. From a sampling frame of several thousand schools, for example, several hundred or several dozen could be drawn. Baseline information on both the types of programs under way and on student outcomes should be measured. Following this measurement, a framework for an observational study should be developed that follows the selected schools over time and monitors their programs and student outcomes. An MDS should be developed as part of this survey that details the types of information that will be required from all sample schools in order to conduct a comparable analysis across them.

- ***Schedule a planning stage for a selected number of randomized trials.*** Schedule in the future (e.g., in two years' time) a planning stage for a select number of randomized trials that will test the most important program variations that have emerged in the early years of the Act. Using the information from the second and third steps above, select a small number of key components and run small-scale randomized trials to measure their net impacts.

II. Conceptual and Policy Issues

The nature and intention of the Act itself poses difficulties in defining the outcomes of interest in any evaluation. To a large degree, the Act is intended to stimulate a major reorganization of existing programs and services and to strengthen ties with organizations defined as partners. A number of outcomes of this type could, in principle, be measured-for example, expansions in the number of students or student-hours served, the number of schools involving partners of various kinds, and so on. To the extent that these outcomes can be defined and measured, this could constitute one type of net impact analysis. The degree to which a particular reorganization is successful will be difficult to compare to the pre-Act organization; all States will be implementing the Act; hence, there is no comparison group. Comparing post-Act organization structures to pre-Act organization structures constitutes a before-and-after design, which has well-known threats to valid inference because the general environment may change in all States over time. One may therefore be left with comparing different types of post-Act organizational structures to each other, although presumably each is relative to its pre-Act baseline as well.

Nevertheless, the ultimate goal of the Act is to improve student outcomes, and the organizational restructuring initiated by the Act is only a mediating variable in the process. Therefore, a complete evaluation must clearly involve the examination of the effects of the Act on students, even if it also includes a study of organizational structures.

The study of the effects of individual types of school-to-work programs on student outcomes must clearly be part of the evaluation. As already noted, there will be dozens if not hundreds of different types of programs, an issue to be considered in the next section of the paper. But suffice it to say here that there are several key hypotheses of the school-to-work approach that should also be considered as goals for testing in an evaluation. Three of these are: (1) that context-based learning, and instructional strategies that emphasize it, are more beneficial for long-run student outcomes than traditional instructional strategies, even for college-bound youth, (2) that one-track schooling is superior to multiple-track schooling for all groups concerned, and (3) that the involvement of employers and other partners in the school-to-work program improves student outcomes.

Having stated that the fundamentals as well as the specifics of school-to-work programs should be examined, it must also be said that it remains unclear whether an evaluation of the Act should be led toward more or less conventional evaluation of the school-to-work programs that spring up under the Act. Many, if not most, of the programs will have been in place in some form in some locations for considerable periods of time prior to the Act, and many have undergone past evaluation efforts (Stern et al., 1995). The Act may have its main effects merely in extending existing programs to greater portions of the student population or in providing them on a larger scale than they were provided prior to the Act. The direct effect of the Act should, in this case, be measured not by studying the effect of any single program on individual students but by studying the effect of the extension of the scale of existing programs-that is, by the incremental gains achieved by bringing greater numbers of students, possibly different in characteristics, into the system than was the case prior to the Act. This type of evaluation is relatively uncommon in the history of program evaluation efforts and raises design problems rarely encountered.

III. Design Issues

There is extensive literature on many of the key elements of an evaluation design (Rossi and Freeman, 1993). Among these are defining the outcomes of interest; defining the treatment(s); picking the unit of analysis; devising a monitoring methodology; and considering process analysis, design of a data collection protocol and, most importantly, choice of a comparison methodology by which the net impact of the treatment is to be measured. In this section, four issues will be discussed: (1) defining the treatment, (2) choosing a comparison methodology, (3) choosing the unit of analysis, and (4) designing a data collection protocol.

1. Program Typology

A proper evaluation requires a detailed typology of program types. While this requirement is obvious in general, the diversity of programs that will exist under the STWOA makes such a typology especially important. A sample traditional program typology is given in Exhibit 1. The list of program types is not exhaustive, but provides a typical list of types of school-to-work programs at

BEST COPY AVAILABLE

a general level. Within each type of program, there are many subtypes of programs as well, with different types of characteristics.

There is a serious difficulty with the traditional typology, however. With it, programs are classified according to general type and not by the underlying characteristics that make them different from each other. Instead, each is treated as a "black box," to use the terminology in the program evaluation literature. With a black box typology, the best an evaluation can do is identify which programs work and which do not work and the magnitudes of the net impacts of those that do work. The net impact estimates of this type provide little information to the policymaker who wishes to improve the features of the unsuccessful programs. They also provide no information on whether extending a successful program to a new location or to a new population, which requires at least one change in a feature of the program, will yield the same impact as the program tested. Traditional black box treatment typologies are generally uninformative on which features work and which do not work.

An alternative typology instead classifies programs by their features, attributes, characteristics, or components. An example of such an *attribute-based* or *component-based* typology is given in Exhibit 2. The first four attributes in the table (whether work experience is paid, whether employers provide financial support, whether there is structured learning in the school, and whether placement assistance is provided) are drawn from Stern et al. (1995, Figure 1) and are examples of a subset of a larger set of important attributes. As Stern et al. demonstrate, the traditional programs in Exhibit 1 differ in the attributes in Exhibit 2 and can be characterized by them. A design matrix for a program evaluation can make use of these attributes to create what is known as a factorial design, which posits programs with different combinations of the attributes in the table. The goal of the evaluation then becomes to determine the effect of each attribute alone, holding the others constant at some level. Net impacts of the attributes themselves would provide much greater information to policymakers who can consider adding or subtracting program features when modifying them.

Realistically, there will be many attributes that will be heavily, if not uniquely, associated with a particular traditional program of the type shown in Exhibit 1. Consequently, a realistic typology will

Exhibit 1

Sample Program Typology: Traditional

Apprenticeship

co-op

Career Academies

Tech Prep

School-Based
Enterprise

School-To-
Apprenticeship

Exhibit 2

Sample Program Typology: Attribute Based

Work Experience Is Paid

Employers Provide Financial Support

Structured Learning in School

Placement Assistance

Scale (Number of Students)

Relative Scale (Participation Rate)

Expenditure Per Enrollee

Expenditure Per Eligible

have to combine the traditional as well as components-based forms. However, there are many components that are common and the goal should be to maximize their use. In addition, even within types of traditional programs shown in Exhibit 1, there are many components- or attribute-based variations that could be built into the typology.

The last four attributes in Exhibit 2 (two scale variables and two expenditure variables) are unfortunately rarely explicitly included in program evaluations despite their deep importance. A traditional classification of programs such as that in Exhibit 1 misses completely the effects of scale and of intensity (of which expenditure is one measure) that cause programs of the same type to have very different impacts. When moving to an attribute-based design, scale and intensity measures become more natural features of the treatment design and program typology. These last four attributes are particularly important for the STWOA for the reasons discussed earlier-namely, that one goal of the evaluation, if not the major goal, may be to determine the effect of an expansion of the scale and intensity of existing programs. In Exhibit 2, this is naturally considered to be a program attribute that is altered by the Act, and therefore its net impact will be a natural outcome of the evaluation.

Finally, an alternative method of building up a program **typology** is to work from a time ***diary*** approach. In this approach, information is collected on how students actually spend their time over the course of a typical school week. The number of hours spent in traditional classes, the number of hours spent in specific school-to-work programs, and the number of hours spent with employers can be gathered. Programs can then be characterized or classified by how they impact students' actual lives and experiences. Whether they have any contact at all with employers and what type of instruction they actually receive, for example, would be a natural part of a classification scheme generated in this way. The intensity of a program can also be more easily measured by the number of hours spent in the activity per week.

2. Choice of Comparison Methodology

A major decision in choice of comparison methodology is whether to use randomized trials or observationalanalysis. Exhibit 3 provides a list of the strengths and weaknesses of each type. There is

Exhibit 3

Strengths and Weaknesses of Randomized Trials and Observational Analysis

Randomized Trials

Strengths

1. Strong internal validity when properly conducted
2. Credibility
3. Relatively simple to communicate to policymakers
4. Relatively inexpensive

Weaknesses

1. Difficulties of field implementation if operators resist
2. Ethical or legal barriers to denial of services
3. Limitation on number of treatments feasible to test
4. Danger of limited external validity and generalizability
5. Difficult to incorporate scale effects, entry effects, and participation effects
6. Risk of a noninformative evaluation; black box problem

Observational Analysis

Strengths

1. Relatively easy to capture wide variety of treatments
2. External validity and generalizability strong
3. Few ethical or legal barriers, or field implementation problems
4. Can incorporate scale effects, entry effects, and participation effects

Weaknesses

1. Risk of weaknesses in internal validity
2. More expensive if properly conducted
3. Data needs are very demanding
4. Can be relatively difficult to explain to policymakers

extensive literature on this issue, and Exhibit 3 provides only the most important elements. (See Burtless, 1995, and Heckman and Smith, 1995, for two recent contributions.)

Randomization, if conducted and implemented successfully, has the major advantage of eliminating threats to external validity (or selection bias). This is by far the greatest strength of randomization and the source of its greatest appeal. Because of this strength, randomized trials have an element of credibility in the research community and with policymakers that is often lacking in observational net impact studies, although one may argue that only weak observational studies lack such credibility. The simplicity of the random assignment methodology also makes experimental results easier to communicate to policymakers than some observational studies, although it may be difficult to separate credibility from simplicity, since even observational net impacts can be presented in a simple table. Finally, contrary to some arguments in the literature on these issues, random assignment methods are generally less expensive than proper observational methods. The lower cost results from: (1) a reduced necessity for collection of preprogram and postprogram data on participants and the environment, (2) a reduced necessity to collect information on the characteristics of the treatment, since in experimental studies those characteristics are set by the evaluation, and (3) a lack of need to collect information on the determinants of why some students are enrolled and others are not, since this is determined by the coin flip. However, each of these cost advantages reflects a potential weakness of the random assignment methodology as well.

Resistance of field operators to implementation of the design and to delegating control of the treatment to the evaluators can be a serious problem. A case in point is the National Job Training Partnership Act (JTPA) Study's failure to secure the participation of the majority of local JTPA agencies which may have affected the generalizability of the evaluation (Hotz, 1992; Heckman and Smith, 1995). Hotz (1992) points out that the problem of operator resistance arises not so much in an experimental evaluation of a new program that is being tested on a small scale, but more so in a nationally based, ongoing program. The school-to-work initiative is of this type.

Randomized trials also suffer well-known resistance to denial of services to the control group, either for ethical or legal reasons. This has been a problem in many experiments but also has been overcome by appeals to the scientific validity of the method and to principles of fairness in the face of limited resources. It can also be overcome by not denying any students a program but by randomly assigning different programs to all students. However, this assigning prevents a study of the overall impact of a school-to-work program because it limits the net impact only to relative comparisons.

Several other weaknesses of the experimental method reside more deeply in its statistical design. Most randomized trials in ongoing programs are based on the traditional typology given in Exhibit 1. Consequently, each program tested is in reality a package of program elements that is often merely rearranged or differently emphasized in other programs. Although a comparison of net impacts of the traditional type can be used to infer the effects of specific program attributes, this requires that a sufficient number and variety of programs be tested to identify the effects of those attributes. In the case of school-to-work, the great number and diversity of programs make this approach difficult. If, instead, only a relatively small number of programs were made subject to experimental tests, the effects of individual attributes could not be identified and the external validity and generalizability of the results would be damaged; policymakers could not use the results to predict the effects of new programs that are slightly different from those tested. Randomized trials are also ill-suited to the estimation of scale effects, entry effects, and program-participation effects (Moffitt, 1992). Typical randomizations occur at the point of program entry or later in the process. In such trials, it is not possible to estimate the effects of the scale of the program because scale cannot be made a feature of the treatment; scale is determined instead by design considerations of how many experimentals and controls to enroll in the experiment. While it should be possible in principle to make the number of experimentals enrolled an experimentally varied feature of the treatment, this requires a further expansion of the number of programs tested, which has already been noted is a problem and has rarely been done in past evaluations.

In addition, and possibly more serious, even varying the scale of the program experimentally does not address the issue of how scale is determined in actual field operation of a program, where scale is determined by a combination of the volume of student volunteers and of the number of slots available (which is in turn dependent on the funds available, the numbers of cooperating partners, and so on). If the object of the evaluation is to determine the magnitude of the effects of the expansions of scale and intensity that States will actually be effecting because of the Act, one must determine exactly how States effect that expansion and what types of new students will be involved in the expansion. It is difficult for experiments to yield information on this question.

Observational studies, which base inference on a comparison of outcomes for different programs in different areas, more easily incorporate information on a large variety of program types. External validity and generalizability are correspondingly strong. Observational studies also have the capability for the study of scale effects, entry effects, and participation-rate effects, although many observational studies unfortunately ignore these issues. Cross-sectional comparisons of outcomes across programs can in principle incorporate information on scale and intensity, program entry, and on what types of students participate and what types do not.

The corresponding weaknesses of observational studies are the greater danger of selection bias (lack of internal validity) and the greater cost of data and information collection. The two are more closely related than they appear. The presence of selection bias arises when programs are compared in which student enrollees are noncomparable and different even in the absence of program differences (econometric discussions of this problem can be found in Heckman and Robb [1985] and Moffitt [1991]). Eliminating or reducing those differences in student populations requires extensive data collection of two kinds. First, information on the students themselves must be collected (panel data histories, for example). Second, information on the reasons for differences in scale and intensity across areas must be determined, since scale and intensity are key features of the treatment and of the Act. This requires gathering information on the programmatic determinants of why certain students end up in a program and others do not, and why that differs across areas. The. ***determinants of the program participation rate*** must be analyzed.

Depending upon the type of comparison to be made, preprogram data for students in programs prior to the Act must also be collected as part of an observational study. If the goal of the Act is to expand scale and intensity relative to their levels prior to the Act, the pre-Act levels must be measured.

The formal data requirements for a proper observational study are far more demanding than is typically realized. Any observational study must specify, implicitly or explicitly, the set of data elements that must be collected from all sites involved in the evaluation to make the determinations necessary not only to measure outcomes and treatments but also to reduce threats to internal validity. The set of data elements is sometimes generically termed an MDS in recognition of the fact that the collection of additional data is allowed. A prototype MDS showing a few representative elements is given in Exhibit 4. The elements listed in an MDS are those that must be collected from every site involved in the evaluation. A listing of the sources from which the data can be obtained must also be included with the listing of the variables in an MDS.

Another reason for the higher cost and data collection needs of an observational strategy follows from the need for many more school observations than are necessary in a randomized trial. In traditional experiments, a few different types of programs are tested in a relatively small number of sites. This traditional approach is handicapped by its black box character-each experiment is a bundle of different attributes. Since there are typically more attributes than programs tested, it is not possible to isolate the net impact of any single attribute, holding others fixed. However, this is not necessary to the experimental methodology and, in fact, many of the large-scale experiments of the 1970's (e.g., the negative income tax experiments) were not of this type because they were not black-box in nature. Randomized trials in the school-to-work area would be best designed to vary by only a small number of attributes-only the most important ones-and to make a goal of the evaluation the isolation of the impacts of individual attributes.

Nevertheless, for cost reasons, it is not possible to conduct more than a handful of such evaluations. For an observational study, on the other hand, collecting data on only a handful of schools would result in an evaluation with too little treatment variation and insufficient power of analysis because site effects-that is, noncomparable differences across schools—would endanger the net impact estimates that are made from across-school comparisons. Gaining sufficient statistical power therefore requires a large number of schools (e.g., several hundred or several dozen) with sufficient treatment variation to enable the reliable estimation of net impact.

The larger number of schools necessary in an observational study is also necessary to make possible a change in the unit of observation from a within-school study to an across-school study. In the former and more traditional analysis, a comparison of the outcomes of participants to nonparticipants within schools is conducted, and the variation in program across schools allows the estimation of the effects of multiple program types. This method is also the method underlying experimental evaluation. In the case of an across-school analysis, however, mean outcomes of all students in a school-both participant and nonparticipant combined-are compared across schools, schools that offer different types of programs. As part of this comparison, of course, controls for

Exhibit 4

Minimum Dataset Prototype

Outcome Variables

Test Scores

Graduation Rates

Employment Rates

Earnings

Treatment Variables

Program Typology

Enrollment Rules

Selection Criteria for Enrollment

Rules for Determination of Number of Slots

Program Information
(Enrollees Only)

Dates of Entry Into Program

Treatments Received

Duration

Exhibit 4(2)

Student Characteristics

Current

History

Family Background

School Characteristics

Local Area Characteristics

Unemployment Rate

Industrial Composition

Alternative Training Opportunities

Note: All individual-level variables are required for both enrollees and nonenrollees except where otherwise noted.

the types of students in the school, the type of local labor market, and other factors are taken into account. However, the across-school comparison is important inasmuch as it provides a test for whether selection bias is present in the within-school analysis. If, for example, a within-school analysis falsely shows a negative effect of a school-to-work program merely because the lower-performing students are selected into the program, this information can be detected by a comparison across different schools with the same types of student populations but different programs or with no school-to-work program. For example, a school with the same type of student population but with no school-to-work program will have the ***same*** mean student outcome as the school with the school-to-work program if the negative impact estimate in the latter school (obtained from the participant/nonparticipant comparison) is false and a result only of self-selection. On the other hand, for example, if a proper within-school evaluation is conducted that uses a valid comparison group, and if the net impact shows a positive effect of the school-to-work program, this impact should also show up in the across-school comparison: the school with the school-to-work program should have a higher mean student outcome than the school with no school-to-work program, because the school-to-work school or some fraction of its students has had its outcomes improved. This method of testing the validity of an observational study and of testing for the presence of selection bias is closely related to econometric methods of selection bias adjustment and the method of instrumental variables (Moffitt, forthcoming).

Although the relative strengths and weaknesses of experimental and observational approaches can thus be delineated, the choice must depend on the program evaluation in hand. In the case of school-to-work, the ethical and legal difficulties in denying services to students with an ongoing problem might be severe. However, if these difficulties can be overcome, the great virtue of internal validity of experimental trials suggests that at least a few randomized trials be implemented in selected areas and for selected program types. Provided local resistance could be overcome and randomized trials could be implemented and their integrity not compromised, they would provide a source of credible net impact analysis, even though limited in scope and generalizability.

At the same time, the limitation of randomized trials, especially for measuring scale and intensity effects, calls for the implementation of a large-scale observational evaluation effort as well. An integrated net impact analysis that combines the results of selected randomized trials with a larger, enveloping observational analysis of a larger set of program types and of scale and intensity effects of the Act could yield a comprehensive evaluation that is superior to either type of approach taken alone.

Finally, it should be noted that a process analysis to not only identify success at implementation but also to identify unique difficulties experienced by individual programs is an essential part of any type of evaluation. A major defect of most past process analyses has been a failure to integrate their results into the net impact, or outcome, analysis. To do so requires a formal characterization of process and the development of measurable indicators of implementation success. These measurable indicators must then be used in the statistical analysis of outcomes.

IV. Implementation

The discussion of design issues in the previous section makes clear that the data collection process, treatment-type delineation, and development of experimental design need to begin immediately for an evaluation of the school-to-work initiative to be successful. The timing of measurement of outcomes must be determined when sites are selected for the evaluation and the timing of their initiatives is determined.

The history of the evaluation of the Job Training Partnership Act suggests that incentives to induce the school-to-work community to cooperate in the evaluation effort will be required. However, it is questionable whether reliance purely on financial incentives will be sufficient or even desirable. Not only do financial incentives affect the nature of the relationship between the evaluator and those evaluated, the provision of significant funds to program sites can alter the estimated impact of the treatment by changing the financial position of the recipients. Financial incentives should be limited to the costs of compliance with the evaluation. Ideally, legislative requirements for participation in the evaluation are desirable, although these are rarely provided.

V. Conclusions

Evaluating the School-to-Work Opportunities Act presents challenges not present in many past evaluations. The virtues of the legislation are in its provision of flexibility to the States to seek the types of programs that fit best their own needs, rather than a Federal mandate or regulatory approach to local programs. This creates difficulties for evaluation because of the variety of programs thereby created, as well as conceptual difficulties in defining the treatment. The approach recommended here is an integrated combination of selected randomized trials of carefully chosen school-to-work programs and of an overarching observational evaluation designed to yield a comprehensive net impact analysis incorporating all important sets of influences and factors.

A four-stage approach to the evaluation was outlined in the introduction. The key aspects of that approach are (1) the development of a component-based program typology; (2) an extensive data collection effort to learn what is going on-how many programs of different types are in operation across the country; (3) a major effort to develop a database for an observational study, involving a large number of schools and students, with an MDS protocol developed with it; and (4) the design and implementation of a few randomized experiments starting in about two years' time.

Net Impact Evaluation of School-to-Work: Contending Expectations

Hillard Pouncy
Center for Community Partnerships
University of Pennsylvania

Robinson G. Hollister
Joseph Wharton Professor of Economics
Swarthmore College

I. Introduction and Background

The School-to-Work Opportunities Act of 1994 (STWOA) provides state- and local-level implementors wide discretion within broad Federal guidelines. It targets several problems in the noncollege youth labor market-too many high school dropouts disconnected from the workforce, a changing workplace challenged by heightened international competition-but it is mainly concerned with the growth over the past decade in college and noncollege skill, earnings, and employment gaps. In response, it seeks a system that connects noncollege-bound youth to career training or higher education in hopes such a system will decrease college/noncollege earnings, skill, and employment gaps. Specifically, the STWOA envisions changing the American education system in ways that will help "youth acquire the knowledge, skills, abilities, and information about and access to the labor market necessary to make an effective transition from school to career-oriented work or to further education and training" (Section 2(5)).

To help achieve these goals and targets, the Act also provides incentives—e.g., planning grants, implementation and development grants, technical assistance-for local partnerships to create the specific programs it recommends.

In this paper, we focus on some guidelines that are over-broad, unclear, or inconsistent and therefore create problems for a net impact evaluation. The guidelines in some cases keep alive contending expectations about what the STWOA is and is not, what local partnerships should or should not do, and what school-to-career programs do and do not do. Local variation is a valued component of the

STWOA's implementation design; but with such wide discretion, local groups may literally implement different theories of a common Federal program (Weiss, 1995). We have selected a sample of key, representative issues that illustrate this and other evaluation problems. The eight issues we discuss include program participation, access and high standards, program leadership, other employer hiring and promotion issues, job market differences, mentoring, learning in context/work-based learning, and access by nonschool and dropout youth.

To deal with the program's lack of clarity, we recommend that some elements of theory-based evaluation strategy be added to a mixed experimental/quasi-experimental evaluation design. A theory-based component may help track how the STWOA practitioners manage the program's broad discretionary features and unclear guidelines.

II. Conceptual and Policy Issues

1. Program Elements

Evaluators first face a problem finding typologies that capture how local partnerships match the program's many elements, like co-ops and youth apprenticeships, with the program's many required activities, like integrating school and work-based learning and connecting academic learning to occupational learning. Glover and King refer to this as a standardization problem-how to provide a standard language with which evaluators may capture the myriad choices local partnerships make.

This standardization problem arises because the Act encourages implementors to "build on and advance a range of promising school-to-work activities, such as Tech Prep, career academies, school-to-apprenticeship programs, cooperative education, youth apprenticeship, school-sponsored enterprises, business-education compacts, and promising strategies that assist school dropouts, that can be developed into programs funded under this Act" (Section 3(a)(8)). It establishes a school-to-work system in which various state and local partnerships use whatever combination of programs suits them to connect schools and work as long as they achieve the five program requirements (Title I, Section 101) and develop an approved governance system (Title II, Sections 201-207).

Of the five program requirements, the key one is that local partnerships must implement the core school-to-work components: integrate school and work-based learning, integrate academic and occupational learning, and establish linkages to secondary and postsecondary education. The other program requirements include the following:

- Provide all students with equal access.
- Provide school-based learning, work-based learning, and connecting activities.
- Provide youth with an understanding of all aspects of the industry they plan to enter.
- Provide career majors.

To secure STWOA state development grants, Governors had to show a governance structure for planning, developing, and implementing that includes nine key state agencies or officials and private sector representatives.

The details of implementation-for example, should employers pay interns or not, should employers help develop curricula or not, should the program target at-risk youth or all youth, should the state adopt youth apprenticeship programs or career academies-are all left to state and local partnerships to decide.

On paper, the STWOA resembles an "administrative sandwich" in which Federal governance guidelines (Exhibit 1, Level 1) and five system requirements (Level 4) bracket specific state- and local-level programs and program components (Levels 2 and 3). In practice, few programs include even the essential components-work-based learning, integrated academicandvocational curriculum, and formal linkages to postsecondary education (see Exhibit 2); fewer still provide all required functions. To carry out the STWOA's core requirements, local partnerships might assemble an assortment of traditional programs that collectively provide essential program components by selecting individual programs for specific roles based on comparative advantages[1] (see Exhibit 3).

---

[1] Cooperative education programs (co-ops), for example, emphasize written training plans and written agreements between employers and youth in ways that link paid employment to schoolwork with supervision and instruction. School-based enterprises provide a work-based learning experience, but usually the work takes place on school premises and students are not paid. Tech Prep programs emphasize formal links to postsecondary education. School-to-apprenticeship programs delay work-based learning until after high school graduation, at which point

Exhibit 1

Level 1

Federal System Level

Federal Guidelines and Funding

Level 2

State and Local Partnership Implementation

Recommended Programs

Tech Prep - career academies - school-to-apprenticeship - cooperative education - youth apprenticeship - school-sponsored enterprise - business education compacts - strategies that assist school dropouts

Level 3

Components and Program Elements

(Employers provide financial support) (Program arranges student work placement) (Work experience is paid)
(Structured work-based learning while in school) (Integrated academic and vocational curriculum)
(Employer involvement in curriculum design) (Formal link to postsecondary education) (Occupational certification) (Pre-elevnth grade career exploration) (Employment/college counseling)
(School curriculum builds on work experience) (Targets at-risk or noncollege bound students)
(Students have mentors from outside school) (Pre-eleventh grade academic preparation)

Level 4

Core Functions and General Program Requirements

Function 1: Integrate school and work-based learning; integrate academic and occupational learning; and establish linkages to secondary and postsecondary education.
Function 2: Provide all students with equal access to the full range of program components.
Function 3: Incorporate three core components-school-based learning, work-based learning, and connecting activities.
Function 4: Provide youth with strong experience in and understanding of all aspects of the industry students are preparing to enter.
Function 5: Provide all students with opportunities to complete a career major.

---

they expect students to enter formal apprenticeships. Youth apprenticeships push up the period of work-based learning and make it part of the high school experience. For its graduates, youth apprenticeships feature links to formal apprenticeships with employers or postsecondary training that include occupational certification. Career academies feature strong curricular features and arrange paid jobs for students related to their field of study; but typically there are no written training agreements or training plans, but they include postsecondary linkages. Their main feature is integrated academic and vocational curriculum.

Exhibit 2

Program Feature	Youth Apprenticeship	Career Academies	co-op	Tech-Prep	School-to-Apprenticeship	School-Based Enterprise	Score
Structured work-based learning while in school	A	R	A	R	S	U	9
Integrated academic and vocational curriculum	U	A	R	U	S	S	9
Formal link to postsecondary education	U	S	R	A	S	R	7

Common features of the six main types of school-to-work programs, sorted by frequency. Programs sorted by number of features included (data from Stern, 1993). Approximate relative frequency of features in school-to-work programs: A=always (score=3), U=usually (score=2), S=sometimes (score=1), R=rarely (score=O).

Alternatively, local partnerships might create more generic programs that at a minimum include the three core school-to-work elements: work-based learning, integrated academic and vocational curriculum, and formal linkages to postsecondary education. Glover and King also suggest that local partnerships might want to differentiate among these generic programs in terms of postsecondary-bound versus work-bound destinations for high school graduates. Stern et al. (1995) made a similar effort, differentiating between school and work for work efforts. (Co-ops are an example of the former, youth apprenticeships an example of the latter.)

Whatever local partnerships choose to do-combine many different traditional programs or make traditional programs more inclusive-evaluators will have a problem comparing program elements across sites. How do local partnerships define participation? Which program components are meant to achieve which functions? Will states and localities have a common language for talking about participation“?

Exhibit 3

Program Feature	Youth Apprenticeship	Career Academies	co-op	Tech-Prep	School-to-Apprenticeship	School-Based Enterprise	Score
Employers provide financial support	A	A	A	R	U	R	11
Program arranges student work placement	U	U	U	R	U	A	11
Work experience is paid	A	U	A	R	U	R	10
Structured work-based learning while in school	A	R	A	R	S	U	9
Integrated academic and vocational curriculum	U	A	R	U	S	S	9
Employer involvement in curriculum design	U	U	S	U	R	S	8
Formal link to postsecondary education	U	S	R	A	S	R	7
Occupational certification	A	R	R	S	A	R	7
Pre-11th grade career exploration	U	U	U	S	R	R	7
Employment/college counseling	S	S	S	U	S	R	6
School curriculum builds on work experience	U	S	U	R	R	S	6
Targets at-risk or noncollege-bound students	S	S	U	R	S	S	6
Students have mentors from outside school	U	U	S	R	S	R	6
Pre-11th grade academic preparation	R	U	S	S	R	R	4
SCORE	28	2s	21	12	10	9	

ommon features of the six main types of school-to-work programs, sorted by inclusion of features. Shaded areas designate a key feature. Programs sorte by number of features included (data from Stern, 1993). Approximate relative frequency of features in school-to-work programs: A=always (score=3), U= usually (score=2), S=sometimes (score=1), R=rarely (score=0).

BEST COPY AVAILABLE

Evaluators must determine how seriously local partnerships follow the STWOA's equal access and high standards requirements. The two goals are potentially antagonistic; and local partners could conceivably select either goal, meet both goals, or avoid both.

Before the STWOA was enacted, some advocates saw the school-to-work movement in egalitarian terms. They expected the STWOA to reduce earnings and employment gaps between college and noncollege youth without creating new gaps among urban/nonurban and minority/nonminority noncollege youth in the process (Robert Crain, Lerman and Pouncy, W.T. Grant Foundation, Joint Center for Political and Economic Studies, National Council of La Raza). Other advocates emphasized the program's meritocratic goals. They focused on college-bound youth and the possibility that the program might blur boundaries between college-bound and career-track programs (Stern et al., 1995).

Supporters of both viewpoints believe that school-to-career systems should close college/noncollege gaps and increase skills; they differ in terms of what they fear from poor implementation. Egalitarians fear a system that reinforces existing earnings and employment gaps between urban/nonurban and minority/nonminorityyouth as it closes earnings and employments gaps between college and noncollege youth. The worst case from a meritocratic viewpoint is this: 'Students who want bachelor's degrees are in some ways more ambitious and are more likely to perform well academically. If these students reject school-to-work programs, the new programs could acquire a second-rate image . . . In the worst case, the programs could come to be seen as another kind of dumping ground for the non-academically inclined, where poor performance and low expectations would reinforce each other" (Stern et al., 1995, pp. 15-16).

The STWOA honors both sets of expectations, and evaluators have to sort out how local partners have solved this equal access/high standards trade off. Specifically, the STWOA requires local partnerships to provide equal access to all youth, regardless of socioeconomic status, limited English proficiency, and so forth (Title I, Section 101(5)). It also demands an uncompromising commitment to high standards. To prevent creaming- restricting access only to elite students-House conferees who wrote the STWOA required local systems to provide equal access. House conferees also made high standards a priority and defined high standards this way: "Students who complete a school-to-work program should be able to enter a postsecondary education or training program, including a four-year college program, without additional academic preparation" (House-Senate Conference Report). Local practitioners are also instructed to ensure the participation of high-achieving youth.

To underscore the seriousness of these twin concerns—equal access and high standards-the Act requires that implementors collect and analyze information on postprogram outcomes "on the basis of socioeconomic status, race, gender, ethnicity, culture, and disability, and on the basis of whether the participants are students with limited-English proficiency, school dropouts, disadvantaged students, or academically talented students;" and look for impacts on the disadvantaged and dropouts as well as the academically talented (Title I, Section 104 (7)).

The Act wants implementors to collect such data because skill differences' often correlate with disadvantage, areas of high poverty, race, and other attributes. If all school-to-work candidates had similar skill levels and only differed by circumstance-disadvantaged, rich, poor, black, white, etc.—meeting the two requirements simultaneously would only be a matter of outreach and inclusion. However, because skill levels differ by circumstance, implementors may choose to emphasize access—making sure low-skilled youth are included-ver high standards; or emphasize high entry and exit skill standards.

Skills, then, are the real issue in this equal access/high standards conundrum. To address this problem, the Act sets aside funds for remedial education and comprehensive support services in

---

[2] Skills are commonly characterized in terms of standard reading, math, and writing assessments by grade level, grade, or test result. Some programs, for example, require 8th-grade skill levels to be admitted to an 11th-grade youth apprenticeship. In its trial year, ProTech, the experimental intern/apprenticeship effort for Boston area hospitals, intentionally set its entrance requirements as low as possible to ensure a large number of participants. They initially required a C+ grade-point average with at least a B average in math, science, and English and a 90-percent attendance rate. Even when ProTech administrators modified standards to a C+ grade-point average with at least a C- average in core subjects and 90-percent attendance, they could fill only 31 percent of the 88 slots they had open. They filled those slots only after they lowered the grade standard to at least a C grade-point average, a C- in the student's last math class, and an 85-percent attendance record (communication with Jobs for the Future). New York City bases part of its admission process to its elite Career Magnet Program on reading test results.

high-poverty rural and urban areas, defined as census tracts with youth poverty rates of twenty percent or more. These areas are likely to have disproportionate numbers of low-skilled youth.

The Act also endorses mentoring and other connecting activities for all partnerships, whether in high-poverty areas or not (Title I, Section 104). These connecting activities also include technical assistance and staff training on implementing program requirements, placement services for graduates, and linkages for participants with other community service.

A surprise in the implementation of the STWOA may be that the access barrier may crop up in suburban as well as urban areas. Early results from the McDonald's Corporation-sponsored school-to-work demonstration led Robert Sheets, a member of the coordinating team designing the curriculum for the corporation, to the surprising belief that average- and low-skilledwhite youth were as locked out of the suburban McDonald's program as their minority and disadvantaged counterparts were locked out of the urban programs. He said that, if anything, the suburban McDonald's site had a higher rejection rate than its urban, high-poverty counterparts, possibly because the suburbs have better alternatives.

Evaluators may also want to pay attention to strategies some advocates already endorse, or practices some sites already try:

- Although some STWOA advocates believe comprehensive services will help increase access, many vocational specialists put their hopes elsewhere. Some, for example, suggest that local partnerships focus on early interventions in the elementary and middle school years that combat skill gaps before they start (Pouncy and Hahn, 1996).

- New York City's solution for balancing equal access and high standards is to mainstream average- and some low-skilled youth by a lottery into its otherwise selective Career Magnet schools. Some, but not all, average- and low-skilled youth gained access to a program for high-skilled youth in a manner that avoided issues of stigma and separation (see Stern et al., 1995). Those average-skilled youth who were admitted by lottery to the selective Career Magnet schools improved their reading scores, attendance, and graduation rates above those of a comparison group of average-skilled youth who went on to regular comprehensive high schools (Crain, Heebner, and Si, 1992). Providing average-skilled youth access to programs requiring high standards may actually help close skills gaps among career-bound youth. Such results have been widely interpreted to suggest that in some cases, bright youngsters do poorly in school because they are poorly motivated and do not see the relevance or "payoff" for effort and high achievement. Their skills improve when a credible opportunity is presented to them. In such cases, providing access to high-standard programs would achieve both the STWOA's goals-equal access and high standards. The strategy does not work for low-skilled youth (Crain, Heebner, and Si, 1992), and how much access can a system provide before it is no longer a high-standards system? Also, some employers have been unwilling to participate in such experiments because they dislike the entrance lottery.

- McDonald's solved the equal-access problem differently. In their program, they only admit youth who meet its entrance criteria, but they provide infinite time and resources for youth from any background to pass its entrance criteria. They provide a preparation course for the entrance exam, and youth who fail the exam can try the course and/or the exam again endlessly. McDonald's strategy avoids a problem Stern et al. (1995) noted in the New York case. Some youth who fail to gain entry to New York's program drop out in disappointment. The McDonald's program provides access for dropouts.

3. Program Leadership

If possible, a net impact evaluation should try to pick up differences in program leadership. Most programs are school-led, but some are employer-led and some are mixed. On paper, the STWOA takes the needs of employers seriously and views them as valued copartners.

> Ultimately a system of work-based learning will need to involve employers more than is currently the case . . . Youth apprenticeship programs will need to evolve over time to make them more consistent with economic and institutional dynamics in this country. Until that time, current initiatives will help to break down the barriers between education and work and provide useful information on program outcomes (Stern et al., 1995, p. 36).

In practice, the real leadership for school-to-career efforts has come from the schooling community and may not serve the narrow self-interests of employers (Whiting and Sayer, 1995).

On paper, the STWOA addresses the leadership/governance issue and requires Governors to assemble a team representing nine interests[3]. In practice, as Gary Walker and Public/Private Ventures noted in a focus-group study of Philadelphia-area employers, an employer-led school-to-career program based upon narrow employer self-interests is unlikely. They found that employers who actually hire recent high school graduates see school-to-career programs as a threat because they believe such programs might force them to share or lose their exclusive access to the limited supply of the area's qualified, work-ready high school graduates.

---

[3] The Governor's group must include the state's educational agency, agency officials responsible for economic development, agency officials responsible for employment, agency officials responsible for job training, agency officials responsible for postsecondary education, agency officials responsible for vocational education, a sex equity coordinator, and other appropriate officials (Sections 203(b)(3) and 213(d)(5), as described by Lauren Jacobs, 1995).

A paper for the Pew Foundation by Whiting and Sayer (1995) picks up on this theme and notes that the employers most likely to become school-to-work partners are those who do not rely upon the high school entry-level market. Even when employers who depend on noncollege youth participate in program development, they do so on a goodwill/community-improvement basis, rather than on a focused, self-interested basis. The result is that employer leadership in this area has been low, leaving the field to school-based reformers.

Recently, more attention has been given to the possibility of employers providing institutional starting points based on narrow self-interests. Here employers define and host the school-to-career effort. The McDonald's demonstration raises the possibility of employer-led implementation. First announced in 1993 (***Implementing Youth Apprenticeship and Related Training Programs for a Comprehensive Career Development Training System in Business Management at Schools and McDonald's Corporation,*** Sheets et al., August 1993), the demonstration is a youth apprenticeship program that differs from other apprenticeship demonstrations mainly in terms of institutional leadership and interest. McDonald's is one of a handful of firms in which Horatio Alger-like, mailroom-to-upper-management mobility still occurs. This policy was reevaluated several years ago when the firm considered switching to college graduates for its management stream. It made a conscious decision to retain its traditional feeder system; as a consequence, its entry-level labor shortages go beyond not finding enough "hamburger flippers." The shortages threaten the firm's ability to replenish its management pool. Secondly, the firm treats urban and inner-city sites as prime locations for its outlets and has, therefore, become a prime urban and inner-city youth employer. It employs twelve percent of all teenage youth (700,000 employees) working in the United States, making it the nation's largest employer of entry-level workers.

In May 1994, the firm launched the demonstration, planning to combine business management classes with work experience for 3,000 students. They designed a four-year program starting in the sophomore year and finishing after one or two years of postsecondary education. After six years of training, graduates manage a McDonald's franchise at a salary of roughly $33,500 a year-not bad for three years out of high school. McDonald's offered to provide school districts with the curricula (developed by Northern Illinois University) and youth with the work experience (twelve to twenty hours) after school and weekends. Certificates were to be earned along the way, from swing manager certification to franchise manager. An initial pilot in two Chicago schools was seen as promising. Other sites include Baltimore; Portland, Oregon; and Muskegon, Michigan. McDonald's goal is to involve 102 schools in fifteen states. Interestingly, firms often find it difficult to coordinate school-to-career programs because they fear poaching and the loss of valued trainees. In this sector, other large consumer service firms face similar shortages and are willing to collaborate-up to a point. McDonald's, for example, will not partner with a direct rival, e.g., Taco Bell, KFC, or Burger King.

We can imagine three governance topologies. One is school led and is generally in effect already. A second is one in which employers organize through a Chamber of Commerce or other consortia, define their interests, and establish a shared leadership position with schooling officials and other public authorities. The local partnership being developed in Fort Worth, Texas, may illustrate this model. A third model features employers who operate in terms of their own narrow self-interests and lead the effort-the McDonald's model. They seek highly skilled youth committed to working for them. These employer-driven efforts mirror school-driven efforts in that, in the long term, these employers hope to induce school systems to partner a school-to-career system on their, the employer's, terms (conversations with Robert Sheets, 1995-96).

Assuming there are enough mixed and employer-led partnerships in the pool of funded programs, a net impact evaluation might assess what difference governance makes. The main idea is that at one extreme, the school-led agenda leaves employers with little incentive to change their intake or contribute much to a new transition system. Why might not the least motivated employer simply select the most highly skilled Tech Prep or other postsecondary graduates and stream them into the firm's slots for college graduates with little additional thought given to restructuring career pathways? That would benefit highly skilled youth, but might it not also reinforce or create new gaps between high- and low-skilled noncollege youth? Implications for youth are that the movement from school to postsecondary school looks very much like a college-bound program, with little thought for youth who cannot afford or see little reason to invest in college.

At the other extreme, an employer-dominated agenda may grant school leaders and teachers too little autonomy. Teachers and school administrators may come to see themselves merely as "parts" suppliers within large corporate webs. The best students and best teachers may not feel challenged enough within such a structure to respond with innovation and attachment.

If a net impact evaluation can include enough sites that capture this variation in governance styles, it may establish whether such effects are real and explore other dimensions of program leadership. An observational evaluation design might be usefully employed to take stock of such differences by governance type.

4. Other Employer Hiring and Promotion Issues

The STWOA provides for the transition of youth from school to work in large terms and in small terms. In its small sense, the STWOA is about occupational certification and formal signaling in the noncollege marketplace. The American youth job market has traditionally been based on informal signaling, meaning jobs are secured by word of mouth with no formal information available from schools to employers and vice versa (Licht, 1992). This has created a situation in which few noncollege youth see a connection between what an employer wants and their own academic preparation. Without a more formal career exploration process, most youth will not hear about employers who create high-wage career pathways for graduates. Similarly, without certification standards, if noncollege young people adjust to school-to-work and do well in school, it will be difficult for them to signal their aptitude to the workplace. The STWOA attempts to eliminate this difficulty by collecting and providing academic information for noncollege youth to employers.

The STWOA could affect labor markets in a larger sense as well. Some fields, like allied health care, for example, have hard credential floors, meaning youth with skills below a certain cutoff point may be barred from a career in the field for which they trained-although in reality, youth with skills below the cutoff point probably are able to carve out some sort of a career. Other fields, like business services, have softer exit-skill requirements, so that youth with average or below average skills may still find jobs within the area for which they trained. Evaluators may want to look at these issues more closely to determine whether exit skills in hard fields reduce college/noncollege skills and earnings gaps, but introduce new gaps among trainees. Similarly, they may want to determine whether college/noncollege gaps are reduced but new gaps crop up in soft fields.

5. Job Market Differences

Since we have never had a national school-to-work system, evaluators may not know what impacts to attribute solely to scale, size, and systemic change, e.g., greater efficiency due to better coordination, less efficiency due to greater complexity. As local partnerships move to scale, do local job market conditions affect the types of system they create? Assuming, for example, that school-to-work programs in markets with plentiful low-skilled jobs differ from markets where such jobs are scarce. are the differences intentional, inadvertent, etc.? Do employers in robust markets who have little trouble finding qualified noncollege labor invest in programs? Or are employers in poor markets more. likely to be conservative?

The STWOA asks local areas to collect much participation information, but not information on the local economy: the types of jobs it offers, levels of growth, unemployment rates, and characteristics of the regional labor market. This type of data is useful for an observational evaluation of what happens to school-to-work at scale.

6. Mentoring

Many vocational specialists consider mentoring an ally of school-to-work programs because it provides missing adult figures for some youth and may have an impact on achievement and career selection. However, Stern et al. (1995) note:

> On an empirical level, the evidence is mixed. There has not been, as yet, a study which conclusively demonstrates the contribution that mentoring programs are thought to be capable of making . . . Mentoring, in this sense, is a "modest intervention." Its power to substitute for missing adult figures is limited. (p. 59)

Perhaps the main value of mentoring is that it helps youth and employers cope with each other's prejudices and attitudes (Pouncy and Mincy, 1994). For example, even when urban minority youth do well in school and an employer can be assured of that fact, surveys suggest that twenty percent of employers seeking entry-level employees will discriminate and reject those youth anyway. The pernicious persistence of racial cuing means that some employers may have more difficulty than expected adapting to a system based on real information about real skills.

Many urban and minority youth do not thrive in informal recruitment networks. As with their suburban counterparts, many fail to invest in academic achievement and flounder through an adjustment period. But some also magnify the reality of employer discrimination and adopt an oppositional affect with regard to all employers and the mainstream labor market, to the extent that they reject excellent employment opportunities when these come along (Bourgois, 1995; Majors, 1992). This combination of employer and youth attitudes is so deadly that for minority and urban youth, the floundering period begins earlier, lasts longer, and is more likely to include a criminal record than is the case for their white suburban counterparts. This makes it even more difficult for minority and disadvantaged youth to climb back into the workforce after they mature. These attitudinal issues on the part of youth present a special challenge to school-to-career programs as well.

Perhaps the attitudes and expectations some minority youth and employers have toward each other can be improved by good mentoring (Anderson, 1990; Bourgois, 1995; Majors, 1992). It is useful for evaluators to determine whether local partnerships see it this way and, in turn, determine if mentoring had such an impact on attitudes.

7. Learning in Context/Work-Based Learning

Some advocates of school reform value work-based learning because they believe it uncaps a different learning mode and provides a new resource for increasing skills. Some workplace reform advocates value work-based learning because they believe it can provide job training relevant to the needs of employers and the offer of high-wage jobs. It not only helps youth to explore career options, but it also provides them with relevant job experience.

In thinking through how to implement the work-based learning component, local partnerships might emphasize its presumed contributions to academic skills. They might implement on employer-based terms and use it to improve job skills, or they might focus on how much is unknown about work-based learning and explore what it does and does not do before adopting it wholeheartedly. Whatever local partnerships choose to do evaluators willwant to determine whether these differences in assumptions are relevant, then assess the impact of work-based learning given what local partnerships sought to achieve.

8. Nonschool Youth and Dropout Access

The STWOA also encourages programs that provide access to nonschool youth and dropouts. One recent program that provides nonschool access and incorporates many of the STWOA's elements is ASAP (**Access**, Supported Advancement and Partnership), sponsored by the Ford Foundation and others. This program and others like it should be included in a net impact evaluation to assess how the nonschool school-to-work component works.

III. Design Issues

A mix of three design strategies may help evaluators address the conceptual issues outlined above. To address the contending expectations built into the STWOA, evaluators might interview local partnerships in depth about their theory of an school-to-work system, what they expect to achieve, and what data they use to monitor progress. Some program components are "black boxes," meaning their effects are unknown or unverified, e.g., work-based learning. Evaluators might experiment with these features to gain a clearer picture of how they operate and what impacts they have. Finally, no one knows what will happen once school-to-work goes to scale. Sorting through how scale itself distorts intentions and intended effects is its own separate category of concern. A large observational study should be used, in part, to help separate effects associated with the program's scale, like increased complexity, from effects specifically attributable to various program components.

1. Finding Out What States Want To Achieve: Theory-Based Evaluation

Common to the eight issues listed in Exhibit 4 is the idea that evaluators may not understand how local partnerships have reframed the STWOA's goals. In such cases, a theory-based evaluation can be helpful for clarifying partnership intentions. A theory-based evaluation is in a sense an extended conversation intended to elicit which assumptions and theories underlie a local program's implementation efforts. The full exercise requires that evaluators determine 'which of the assumptions underlying the program break down, where they break down, and which of the several theories underlying the program are best supported by the evidence" (Weiss, 1995).

Exhibit 4

Issues	Limited Theory- Based Evaluation	Small Experiments	Large Observation	Priority
A. Elements	Applies	Applies	Applies	Early
B. Equal Access	Applies	Applies	Applies	Early
C. Governance	Applies		Applies	Middle
D. Exit Skills	Applies		Applies	Late
E. Market Data			Applies	Late
F. Mentoring	Applies	Applies	Applies	Middle
G. Work-Based Learning	Applies	Applies	Applies	Early
H. Nonschool	Applies		Applies	Middle

We recommend a limited version of this evaluation technique be used for several key issues. For example, how do local partnerships detail the means by which their youth apprenticeships, coops, and other programs will achieve the STWOA's core functions? Does the local system they envision embody meritocratic or egalitarian expectations, both, or neither? How do they perceive issues related to exit skills? Is the work-based learning component exclusively academic, or does it also provide relevant job training? How do they provide nonschool access to their system?

Selected local partnerships should "specify their program's theories and outcomes with as much clarity as possible; specify the proximate or intermediate outcomes as accurately as possible and link each proximate outcome to the program's theory; and specify long-term outcomes and relevant data linking both to theory and proximate outcomes" (Weiss, 1995).

As a more detailed picture of local partnership choices emerges, evaluators may discuss the collective impact of local program variation.

2. Opening Black Boxes With Experiments

In several cases, no one knows whether a critical program component works or not. Evaluators want to know more about those components, how well they work, why they work. The four issues that present such special problems include the impact of specific types of programs as modified locally, the equal access question, the impact of mentoring, and the specific impact of work-based learning.

To determine, for example, whether mentoring can affect the attitudes of youth and employers, evaluators might select a site that brings youth likely to have oppositional attitudes into contact with employers likely to discriminate racially. At that site, evaluators might randomly assign supportive mentors and mediators to a sample of students.

To assess the impact of work-based learning, evaluators might select a school program that matches the Office of Technology Assessment's (1995, p. 5) several criteria for a strong program. Assuming the program is oversubscribed, evaluators might ask administrators to run a parallel program without a work-based learning component. The site should include a significant number of high -skilled and college-bound youth who apply for the courses containing the work-based learning component to test the STWOA's assumption that some high-achieving youth may learn better and retain more when they learn in context, rather than in the abstract.

Evaluators might also experiment with other key program components, such as the salience of a link to postsecondary schools. To test the salience of links from high school programs to postsecondary education, for example, evaluators might look for an oversubscribed youth apprenticeship program that does not include such links. They might encourage administrators to establish a parallel version at a local community college and then randomly assign youth to the two programs.

Evaluators might also try to replicate Robert Crain's experiments on the impact of equal access on low-skilled youth when they find other equal-access, natural experiments or lotteries.

None of these experiments can control effectively for other, unknowable background conditions (Hollister and Hill, 1995), but they can help clear up mysteries about critical program components and partially assess how well they work.

3. Large Observational Study Assessing Systemic Change

Although evaluation specialists agree that observational studies for programs as large and diffuse as school-to-work cannot eliminate selection bias (or the effort to do so is prohibitively expensive (Heckman)) ultimately, they agree that a large, observational study is on balance the more "satisficing" means of assessing the net impact evaluation's core issues: the STWOA decreased floundering, closed earnings and employment gaps, and introduced a new system for connecting youth to jobs.

Specialists also suggest that such a study run for up to 6 years, 1996-2002, with data from as many sites affected by the STWOA as possible. These data might include governance differences, standardized descriptions of local program components and features, time series data on local markets, occupational certification results (exit skills), data from nonschool programs, as well as time series data required from funded programs and time series survey data of local school-to-work partnerships collected by Mathematica Policy Research, Inc.'s implementation study currently underway. Evaluators should also try to construct modalities that capture the patterns and tendencies in programs implemented by local partnerships.

We add that some issues are related to large system change. One might imagine, for example, that an employer-led system might look different, operate differently, and have different impacts for employers than a school-led system. Similarly, if exit skills matter, then they, too, should affect system character. The characteristics of local labor markets might shape a local system, and eventually, shouldn't a local program affect the local market? If mentoring works, then at scale, that too should change the character of minority opportunity. We believe an observational design has advantages for addressing system-changing questions and issues.

4. Priorities

Congress wants to know what the STWOA's impacts were by 1998. Evaluation specialists suggest providing as much of a net impact assessment as possible by then, but they also recommend that evaluators request extensions to complete a larger study by 2002. It is also the case that the school-to-work idea arrived swiftly and it may depart swiftly. Congress is considering terminating the STWOA and folding its funds into an omnibus training and education block grant. The most useful part of a STWOA evaluation may be providing as much information as quickly as possible to the Nation's lawmakers about the STWOA's critical parts: the effect of work-based learning and key components on high-achieving youth, how well equal access works for low-skilled youth, and whether some forms of mentoring help bridge gaps between some minority youth and employers.

IV. Implementation Issues

1. Four-Part Net Impact Evaluation

Given that no evaluation specialist can imagine a single, perfect net impact evaluation instrument or design, perhaps there is strength in numbers-in two senses. For some program impacts, a mixed evaluation design may capture consistency of impact across different evaluation designs-for example, if the postsecondary education component consistently associates with high-wage, entry-level jobs in data from the implementation study, random trials, and the observational study. That pattern should strengthen confidence in each individual study's results. Interview data should then tell evaluators whether the association is a surprise to local practitioners and is an unintended consequence, or it is no surprise and the result is intentional.

A mixed design may also capture additive effects. For example, the Mathematica implementation study and a limited theory-based evaluation may both yield typologies about local programs and their components that might then be added to an observational study to look for impacts by a variable summarizing the various typologies. Alternatively, information on how work-based learning works from random trials might be added to an observational study to develop hypotheses for the larger study. To carry out such a mixed design, we recommend a four-part evaluation implementation (see Exhibit 5).

Exhibit 5

Issue	Implementation Study	Theory-Based Part	Random Trials	Observational Study
Elements	Applies	Applies	Applies	Applies
Access	Applies	Applies	Applies	Applies
Governance	Applies	Applies		Applies
Exit Skills	Applies	Applies		Applies
Mentoring	Applies	Applies	Applies	Applies
Market		Applies		Applies
Nonschool	Applies	Applies		Applies
Work-Based Learning	Applies	Applies	Applies	Applies

Part One

The Mathematica implementation study has been underway since fall 1995. The data they are collectingwill be used to assess the composition of partnerships, program designs, specific activities, linkages to postsecondary options, approaches to student assessment, levels of student participation, and aggregate measures of transitions out of high school and into postsecondary activities (Mathematica brochure). The data are from three sources, including a three-wave mail survey in fall 1996, 1997, and 1999 to assess the progress of local partnerships in creating local systems. They also will conduct case studies of forty-two local partnerships in spring 1996, 1997, and 1999 to capture promising practices and the nature of participation. The forty-two sites include four local partnerships in each of eight states that received implementation grants plus ten local partnerships that received direct Federal grants. That eight-state sample includes Florida, Kentucky, Massachusetts, Michigan, Maryland, Ohio, Wisconsin, and Oregon. Finally, to measure outcomes, Mathematica randomly selected thirty-two partnerships in these eight states for a three-wave student survey of twelfth graders in spring 1996, 1998, and 2000. They will survey eighty-five students in each partnership and follow up each survey with a phone call eighteen months later to assess postsecondary activities. Mathematica plans to make some of these data available on an annual basis, and a net impact evaluation might begin by building on the patterns and modalities Mathematica may have already picked up.

Part Two

If Mathematica were amenable, net impact evaluators might piggy-back theory-based interviews on Mathematica's spring 1997 wave of forty-two case study site visits. Failing that, such interviews should be done with a comparably sized sample as soon as possible. These interviews should not be restricted to older, established sites, but should include local partnerships that are either still thinking through their program or have just finished that process. Again, the key to the effort is to learn what theories (or lack thereof) lie beneath local program designs. How did they perceive work-based learning working? What roles did employers play'? Were employers participating out of civic spirit, or did they define their participation in narrow, self-interested terms? Did they have views on mentors and discrimination? Given their goals, what data would they use for a self-evaluation? It would be particularly helpful to include some of the McDonald's sites for these types of interviews.

Part Three

Net impact evaluators should conduct small experiments to assess the impacts of such key program components as work-based learning and mentoring.

Part Four

By fall 1997, net impact evaluators should be able to construct program modalities or patterns of implementation that they wish to examine at scale with a large observational study running as long as possible that, at a minimum, would capture a four-year cohort of 1997 high school freshmen and youth who graduated in 1997. Such a study would mainly compare modalities, observing which school-to-career designs seemed to close college/noncollege achievement and employment gaps. Which widened or closed college/noncollege gaps without adding new gaps among noncollege youth? How did local labor market conditions affect program outcomes? Did scale and complexity distort some programs? The design might even include a large panel of school-to-career and college-bound cohorts to be monitored indefinitely.

V. Concluding Section

Evaluating the STWOA and the several contending expectations built into it reminds us of the bank robbers' problem that a philosophy graduate student once used to make a point about why meaning and intention should not be taken for granted.

Two imprisoned bank robbers spend all their time behind bars planning a bank robbery they will commit once they are released. They are meticulous and nothing is left to chance. Their plan is elaborate, but they agree on every detail. After their release, they gather together all the equipment they need, they rehearse the crime and execute the robbery flawlessly. However, a few hours after the robbery, it becomes painfully clear that after years of careful coordination, agreement, and partnership, they live in hopelessly different worlds about why banks should be robbed. One robber is an old-fashioned criminal who likes to take the money and run. The other is a would-be Robin Hood who robs for the revolution. After the robbery, the second robber goes into the town square, hands out his share of the money, and hands himself in to police to publicize the revolution. As a result, the first robber is soon caught as well, and they both return to jail.

The STWOA, too, is unclear about its meaning and intentions to its many partners. The STWOA's partners probably agree on its overall goals-reduce skills and employment gaps between college and noncollege youth. They also probably agree with its method and most of its intermediate guidelines, e.g., create a new school-to-work system, integrate vocational and academic learning, increase the relevance of work-based learning, and increase links to postsecondary training. However, differences in interpretation over how to accomplish these goals and differences about priorities may lead to vastly different outcomes and impacts, some of them unintended. To understand the STWOA's net impacts, evaluators should take account of the many unique decisions and choices its state,, local, and private partners have made. We suggest a four-part evaluation model that includes a theory-based evaluation component to capture the implications of differing expectations about a large new school-to-work system.

Summary of the Roundtable Discussion

I. Introduction

The roundtable met on February 23, 1996 at the Doubletree Inn in Washington, D. C. The discussion was, for the most part, free-flowing, and, generally, did not follow a pre-set agenda. Section II summarizes the proceedings, presenting the salient points made in the discussion in the order in which they were presented, identifying the speakers in each case. Subheads flag those points in the meeting where the discussion tended to focus on a particular subject or issue. Section III summarizes the broad conclusions of the large majority of the participants and lists those issues and areas of concern that individuals particularly urged the Departments to take account of in mounting a net impact evaluation of school-to-work.

II. Summary of Proceedings

At the outset, Deputy Assistant Secretary of Labor Ray Uhalde outlined the charge to the roundtable participants: providing recommendations on if, how and when a net impact study of school-to-work should be conducted, giving particular emphasis to the "how" question. As background for the participants, Nevzer Stacey and David Goodwin outlined the overall evaluation strategy for school-to-work and described the individual research projects that are already underway.

Rebecca Maynard opened the discussion by addressing the question of whether we should evaluate school-to-work. She said that we should recognize that failure to look at school-to-work critically now could result in a large investment being made in reform without knowing whether or not it was a worthwhile investment until twenty to thirty years had elapsed. She stressed that schools perform to expectations and it is important to convey to the schools implementing the program the outcomes that we expect them to achieve.

Marion Pines cautioned that we should recognize that we are involved in basic system reform and that focusing on the limited experience of early participants may be a mistake.

Robert Moffitt asked if, recognizing that considerable school-to-work programming is already in existence, should we be measuring the incremental effects of this initiative? Expansion of program scale and integration of systems have not typically been the subject of traditional evaluation strategies which measure programs compared to no program treatment.

J. D. Hoye expressed support for the concept of evaluation but added the concern that indicators of success be selected that move beyond short-term outcomes and "body counts" and that are appropriate in measuring the building of a new system. Like Pines, she is concerned about an inappropriate evaluation approach possibly damaging the program.

James Heckman said we should recognize that in introducing these new innovations, we may be affecting the entire educational system. Our evaluation, therefore, should be broadly conceived to capture the system-wide, indirect as well as direct effects of these innovations.

Chris King identified a consensus in several of the papers that we need a balanced evaluation strategy: a quasi-experimental piece, an implementation piece and some tightly controlled random assignment experiments. He added that the experiments should be focused on those elements of school-to-work that appear to have some stability so that what we measure has some currency.

Marion Pines said one area missing in the papers was the area of curriculum reform -- career majors, career clusters, specifically. This subject should be given a prominent place in the evaluation study.

Rob Hollister pointed out two areas that should be given greater attention. He said that if school-to-work is going to succeed the behavior of employers must change as well as the behavior of schools. Thus, more needs to be done on the employer side. He also asked what we know now about how youth make the transition from school to work. He urged that an evaluation study look for changes in these processes of making the school-to-work transition.

James Heckman suggested we need to step back and examine the premises on which many of these programs are based, e.g. that "floundering" of youths in the labor market is necessarily a bad thing. Before we look at particular programs we should look at what the problem is and how it has changed. He asserted that the factual basis supporting the major premise underlying school-to-work is weak.

Rebecca Maynard advised looking at school-to-work from the school reform perspective as well as the labor market perspective which had been the focus of the roundtable discussion so far. She expressed concern that, while waiting for the results of a long-term evaluation, there should be a short-term feedback loop for schools and employers concerning specific elements that seem to work best.

Hillard Pouncy said we should add to outcome measures whether, in response to school-to-work, employers restructure their career ladders and offer high wage jobs to high school graduates. Another outcome measure would be whether a school-to-work curriculum attracts and retains high-skilled (i.e., high basic skills) youth. He favors disaggregating the "black box" and looking at program elements.

Robert Glover identified a need to focus on outcome measures and bring together both education and skill standards and a need for short-term progress measures for employers and schools as well as long-term net impact results.

Chris King said that overall systemic reform does not lend itself to net impact study. Policymakers need to identify which parts of school-to-work they want to examine and which parts lend themselves to net impact study.

Reid Strieby agreed that we need to focus on specific program elements and stressed the importance, as a practical need for administrators, of short-term progress measures. He added that we will have to show results if we hope to institutionalize those program elements that work.

Marion Pines referred favorably to the ten conceptual/policy issues listed in the Dayton paper and suggested that these should be tested for program outcomes in the study.

Charles Dayton said we should go forward with a study. The ten policy issues listed in his paper will be difficult to measure but they do define what the legislation seeks to accomplish.

Robert Glover also liked the list of issues in Dayton's paper but noted that others might suggest additional issues such as mentoring or career exploration.

Disaggregating the "Black Box"

Recognizing that a consensus was emerging that the evaluation study should disaggregate the "black box" of the total school-to-work initiative, Karen Greene asked the question: How do you disaggregate the components of school-to-work to determine which of these elements is working effectively?

Robert Glover agreed that our approach should emphasize the components of the "black box." In identifying these components, he suggested looking at existing evaluation studies, such as the study of career academies, which found that the "school-within-a-school" concept appears to result in reduced dropout rates.

Charles Dayton noted that bringing new elements to models makes for a stronger program. He cautioned against disassembling effective approaches for evaluation purposes.

Peter Rossi emphasized that, in disaggregating, we look at changes that are actually occurring inside the schools rather than in the ideology of state school systems. He added we should then see how these changes affect outcomes.

Robert Moffitt heard a consensus within the group that the study should get within the "black box" of school-to-work and that how to do it is the crux of the problem. He referred to the typology in his paper which characterizes programs by a specific set of attributes and suggested it can provide the basis for measuring the variation in school-to-work programs. He also noted that a more radical typology, being used in the welfare area, is a typology which is based on what youth actually do, hour by hour, in a typical school day or week. He recommended that the typology be developed at the outset rather than post hoc as was done in the case of welfare employment and training programs.

He also advocated an observational rather than an experimental approach to this evaluation, asserting that the observational approach will allow the evaluators to observe more sites and more variations at a given budget level.

Testing Basic Assumptions

Chris King posed the question of what policymakers hope to learn from this study. J. D. Hoye responded that they wanted to validate certain assumptions that are at the basis of the school-to-work program. Rebecca Maynard said, if this is the case, this argues for a series of small, interim impact evaluations that will test the validity and generalizability of these assumptions concerning what youth need to learn, role of employers, etc. She added that this needs to be distinguished from the broader problem of measuring the impact of school-to-work as a new systemic change in the way we deliver educational services.

Peter Rossi mentioned we need to be clear about what unit we are looking at: schools, school systems, states? If it is states, he does not believe we will be able to go beyond simply providing descriptive data.

J. D. Hoye stressed the importance of looking at the intensity of a given program intervention, i.e. how many hours or days does the student receive a particular program element. She suggested we then try to determine the five or six key cross-cutting elements that are needed to touch every person in the system and make changes in the lives of the people we are trying to affect.

Peter Rossi said, if this is what is needed, one doesn't need a net impact study. One could look at the existing knowledge base on how youth previously made the transition and see whether this new intervention makes any difference.

Responding to a question from Marion Pines regarding which policy assumptions are integral to the school-to-work concept, J. D. Hoye listed the following:

- An instructional strategy should be contextually-based and those who deliver the strategy should have a background in contextually-based instruction.

- A context-based curriculum should be applied to all students, including college-bound, without lowering standards.
- Students and employers both benefit, from a cost-benefit point of view, from bringing students and employers together.
- Exposing students to a wide variety of occupational options through school-to-work will reduce their "churning" in the labor market.
- At a different level, another premise is that a funding strategy of using seed money or venture capital will result in system reform.

James Heckman challenged the assumption that "job shopping" or "churning" by youth in the labor market is necessarily a bad thing. He said that research indicates that there are substantial gains from "job shopping." The question is how far our study should go in investigating some of these basic assumptions. At an even deeper level, can we investigate the economic and political forces that created the modern high school in the United States as a replacement for training by employers under traditional apprenticeship systems? Also, in the area of basic research, we should investigate the extent and effectiveness of private sector training and subsidies to increase private sector training of youth as a policy option.

J. D. Hoye replied that the question is when churning should occur: Is churning more beneficial in secondary school while connected to learning or later after education ends? Related to this, Hillard Pouncy said we should consider the involvement of employers in the curriculum development process and test whether this involvement makes a difference.

Hillard Pouncy stressed that school-to-work presents risks as well as opportunities. One of the risks is that youth with limited basic education skills will not be able to enter certain programs with high entry thresholds. This lack of access could further widen the earnings gap, by race, for example.

Overall Design of Evaluation Strategy

Rebecca Maynard asserted that right now we are building on no information. She asked if there are some short-term studies which would guide system reform -- provide constructive feedback and opportunities to change course. She noted that this gap closing versus gap opening issue is the type of issue that could be addressed in such a study.

Charles Dayton suggested that we rely on the school-to-work implementation study to meet the Congressional mandate for evaluating the broad, systemic educational reform envisioned by the legislation and take advantage of the natural variation among states, to then identify local examples that approximate various models and measure their effects on participants.

Robert Moffitt said we need to begin thinking immediately about a data collection strategy: developing a national baseline school or student sample so that we will know what is occurring right now and be able, in subsequent surveys, to determine the effects of school-to-work.

Rob Hollister noted that a national sample won't allow you to determine the effect of local labor markets on these programs.

Chris King noted that the discussion thus far suggests a three-track evaluation strategy: basic research to get at some of the fundamental labor market issues involved, such as those proposed by **Heckman**; study systemic educational reform through the school-to-work implementation study; and use net impact studies, possibly in a demonstration project setting, to look at individual program interventions or packages of elements.

Case Study Approach

Marion Pines raised the possibility of looking at key assumptions listed by **Hoye** through in-depth case studies. Peter **Rossi** expressed doubts about the case study approach because it does not allow for sufficient comparative analysis, as is possible under such methods as the longitudinal study of cohorts being used in the school-to-work implementation study. Charles Dayton felt this is not an either/or issue. He would like to include in-depth case studies in the evaluation strategy and collect participant outcome data as well.

Identifying Components/Attributes to be Studied

Rob Hollister suggested the implementation study should be used to identify and characterize those modalities that could be tested either through random assignment or comparison groups. He endorsed Moffitt's suggestion that the baseline be established, possibly by expanding the sample in the implementation study.

David Goodwin noted that there might not be enough variation in the sites visited in the implementation study. Peter Rossi responded that this is a problem with purposive samples -- you don't know whether or not the lack of variation is characteristic of the universe. The solution is a representative sample.

Robert Moffitt suggested that using the NLSY sampling frame of 9,500 schools would have two advantages: (a) it would provide a baseline for evaluation and (b) it could supplement school data from the implementation study sample if that sample turns out to have insufficient variation.

Hillard Pouncy recommended inclusion of variables in the study describing the degree of employer involvement in school-to-work. He also raised the possible tradeoff between program quality and access of average- or low-skilled youths to the program and suggested that this issue be included in the implementation study.

Outcome Measures

Peter Rossi suggested we think of a set of proximate outcome measures to initially design the study around and then change the outcome measures as school-to-work matures. Charles Dayton said one of the earliest outcome measures should be attendance. He added another important outcome variable would be whether or not the youth had formulated career plans.

Hillard Pouncy said we should test whether providing information about the job market and the participation of employers increases basic skill levels as measured by standardized tests.

Reid Strieby noted that school-to-work is intended to create a career "corridor" extending from kindergarten through the twelfth grade or later. He suggested measuring the impact of creating career awareness.

Small-Scale Trials to Test Assumptions

Martha Huleatte noted that, as a systemic reform, school-to-work assumes that the program will include college-bound and non-college-bound youth. Marion Pines added that school-to-work designers and practitioners don't want it to have the stigma of being vocational education. She added that everyone goes to work eventually and young people coming out of college frequently don't have a clue about the job market.

Gary Burtless challenged the assumption that school-to-work will benefit equally college-bound and non-college-bound youth. He urged that this proposition be tested. Rebecca Maynard agreed and suggested other assumptions implicit in the legislation that should be tested: that contextual learning curricula result in higher skill achievement levels than traditional abstract learning, and that a school-to-work model is equally effective for students at various levels of academic ability. She noted that these don't have to be long-term tests. One could randomly assign students to one or two years of a new model.

Reid Strieby noted that in schools with both school-to-work and traditional programs, one has a natural experimental setting in which you can test contextual versus noncontextual learning.

Chris King noted that school-to-work grew out of the "forgotten half" discussion and suggested we back up and look at what the reality of school-to-work is rather than idealized models.

Rebecca Maynard distinguished school-to-work evaluation from the National JTPA Study. That study evaluated an in-place program -- title II-A of the Job Training Partnership Act (JTPA) -- whereas school-to-work is being implemented using new ideas. Further, school-to-work is being implemented without a firm knowledge base. She expressed optimism that smaller scale studies -- testing assumptions such as the value of contextual learning --will be more valuable for the process in the case of school-to-work than was the JTPA evaluation.

Peter Rossi agreed with the concept of small-scale trials of basic assumptions, but cautioned that more than one trial will be needed for each assumption/model/variation.

Hillard Pouncy suggested testing hypotheses related to the role of mentoring: does it reduce resistance to learning, does it break down employer discrimination?

Chris King noted that we need a lot more work at the front end of the evaluation. He suggested using focus groups of practitioners/policymakers to identify what are the key interventions and levels of intensity to be tested.

Rob Hollister suggested asking program staff at various levels to identify, in detail, what they feel specific interventions are expected to accomplish. This could be the basis for constructing hypotheses to be tested.

Robert Moffitt felt that investing limited funds in four or five small-scale trials now would be mistake. There will be tremendous variation in school-to-work programs, as Congress intended. We can learn about this natural variation in program approaches through an observational study. He preferred a balanced strategy outlined in his paper which provides for some small-scale trials plus a major study of the natural variation that is occurring. To select four or five components and fund research on only these components would be premature and run the risk of not learning about much that states would want to know.

Gary Burtless disagreed with this approach. He noted that, in a large-scale observational study, each community analyzed will have its own unique configuration and one will not have sufficient statistical power to make meaningful comparisons.

Rebecca Maynard said the reason she presses experimentation on key assumptions now is that a major, costly reform is going forward affecting millions of children with very little information about likely effectiveness. This implies large human capital costs for both the schools and society if the assumptions are wrong. If we wait for the results of a large-scale study, we will be too late -- schools and children will have already made the investment.

Marion Pines raised the political issue of asking the states to make huge institutional investments in school-to-work and then saying we don't know if our basic assumptions are correct. She asserted that there is already a considerable literature to support the law's basic assumptions. Peter Rossi replied that if this is the case then we should commission a "meta-analysis" of this literature before we begin designing small experiments.

James Heckman suggested private sector training as one area which needs further study.

Robert Moffitt pointed out that data collection now to find out what people want to test is necessary to make the proper evaluation design later.

Listing of Research Issues

At this point, Karen Greene listed several specific research issues that had emerged from the discussion thus far and asked the participants for additional issues. The following were the issues listed and those contributed by the group:

- Is school-to-work effective for different groups, e.g., for the college-bound as well as the non-college bound?
- Does employer involvement make a difference?
- How do we track out-of-school youths and how do they fit into the school-to-work system?
- How effective are different instructional strategies, e.g., contextual learning?
- What is the relationship between access and quality within schools and within labor markets?
- Should you pay participants for work performed under the program?
- What is the value of mentoring?
- What amount of time should be spent with employers versus in school?
- What are program costs and is school-to-work cost-efficient?
- What types of employers get involved in school-to-work?

- Does inclusion of organizations such as labor unions and CBOs in the partnership make a difference?

The participants agreed that this was a "laundry list" of issues, some of which were appropriate for an observational study; others for a literature review and others for field testing.

Unit of Observation/Measurement

Robert Moffitt said the unit of observation for these issues should be the school and the school population.

Rebecca Maynard asked how you distinguish participants from non-participants in a school which embraces one aspect of reform school-wide (e.g., contextual learning) but enrolls only some students in another aspect of reform which may have only minimal impact on the rest of the school (e.g., career academies). Gary Burtless replied that the answer to this problem is the approach suggested in his paper in which certain school districts that are not involved in reforms serve as the control observations against which you measure the experience of all youths in those schools that have adopted and are implementing school-to-work programs. Maynard rejoindered that there are very few schools in the country that are not reforming or that will acknowledge that they are not reforming.

Robert Moffitt pointed out that the limited funds for school-to-work will force States to limit the number of schools participating in reforms.

A colloquy developed on the question of whether or not the school as a whole should be the unit of observation. Specifically, the issue was raised of whether spill-over effects of reforms within a school will affect non-participants as well as participants.

Reid Strieby raised the difficulty of knowing what we mean by "implementation" of school-to-work. The degree of implementation may vary greatly among schools even within a school district that has adopted school-to-work reforms.

James Heckman strongly endorsed Robert Moffitt's model of analyzing fundamental attributes of programs and interactions among these attributes since the programs themselves are in transition.

Concluding Comments

Nevzer Stacey, the roundtable discussion leader, then asked each of the participants to offer concluding comments:

Robert Glover agreed with focusing on certain key characteristics of programs.

Rob Hollister expressed pessimism about using a broadened implementation study and the attribute-based typology proposed by Moffitt to study schools., We may not be able to capture the subtleties involved in this typology as compared with more straightforward measures such as expenditures and pupil/teacher ratios. Asked about whether the implementation study sample should be broadened, he suggested that it might be better to await completion of the first round of the implementation study and then decide if there will be sufficient payoff from a broadened sample.

James Heckman concluded it would be useful in our analysis to spell out more precisely the theory underlying these programs and cite the available evidence concerning these theories. This will permit a dimensional reduction which would be valuable if, as Burtless suggests, school-to-work programs will involve a hopeless degree of complexity. This approach will also reduce the need for randomization which would be needed in a program-based approach. Heckman endorsed a sequential analysis in which one first determined what the base characteristics are and then decided whether to look at a few schools intensively or look at a broader range of schools. He believes doing these simultaneously would be unwise.

Robert Moffitt suggested a four-step evaluation strategy:

1. Develop a program typology.

2. Find out what's going on, partly as insurance in case not enough is learned from the currently planned implementation study. One possibility is to "piggy-back" on the fall 1996 NLSY survey to supplement the 64 schools in the implementation study.

3. As soon as possible do a first-round, baseline representative survey to set the stage for an observational analysis.

4. Consider a small number of randomized trials in about two years after you have completed steps 1-3.

Marion Pines concurred with the Moffitt approach. The typology should consist of program components rather than complete programs such as Tech Prep. She would include a school day typology in the evaluation. She also endorsed obtaining validation of some basic assumptions through randomized trials.

Reid Strieby concluded we should start with small-scale trials to validate assumptions.

Charles Dayton said to include in the implementation study an examination of the extent to which the ten policy elements of the Act cited in his paper (or another comparable list) are showing up at the state and local levels. As part of that process, the evaluation should identify a typology of programs and then measure participant effects of these programs, program elements or blending of elements against control or comparison groups. He also urged that the evaluation gather cost per participant data at each site.

Rebecca Maynard concluded we should remember that school-to-work presents a very different context for evaluation than have previous large-scale evaluation efforts. She agreed with Moffitt's four-part approach but felt there is a missing step: we do not know the validity of the assumptions underlying these school-to-work models. These principles should be tested early in the process rather than three or four years down the road.

Gary Burtless suggested a four-step approach to the evaluation:

1. Implementation analysis, including use of management information, to determine the effectiveness of implementation, as "effectiveness" is defined in the legislation.

2. An observational study which would attempt to tease out from the naturally-occurring variation across school districts in the country information about relative program effectiveness in terms of impact on students.

3. Randomized trials to determine the effects of planned variations in program (e.g., adding mentoring to the program) and whether the variation helps to make the program more effective. Ideally, he agrees that it would be desirable to do randomized trials to test basic assumptions but this would involve denying treatment to controls. In the case of planned variations, you have two variants of a treatment so that everyone in a school is receiving some treatment.

4. An employer survey in those communities where school-to-work has reached a significant scale to determine: (a) from employers participating in the program, whether the program is meeting their needs and (b) from non-participating employers, why they are not participating and the changes in the program that would make it more attractive.

If limited resources require setting priorities among these four evaluation components, Burtless' order of preference would be: the implementation study, randomized trials, the employer survey, and lowest priority for observational study. He is very skeptical about the observational study because if, as today's discussion suggests, virtually all school districts are implementing school-to-work, an observational study will have limited power to determine relative effects of program components. Also, an observational study will be dealing with an immature system which will be very different in five years.

Hillard Pouncy stressed the importance of developing a typology by consulting those who implement the program. He would also like an evaluation to focus on the youth who participate in school-to-work.

Peter Rossi concluded it is clear that school-to-work is a chaotic system based on a lot of untested assumptions. This underscores the importance of constructing a typology of attributes and then measuring these attributes through an observational study. He strongly favored a literature review on some of the basic assumptions underlying school-to-work as well as on what we know about the school-to-work transition process itself.

III. Conclusions and Key Issues/Areas of Concern

As noted at the outset of this report, this roundtable discussion was not structured to force agreement or consensus among the participants. Yet, as the debate unfolded, areas of broad general agreement did emerge, key issues were sharpened and many important individual areas of concern were identified. This section draws general conclusions from the discussion and identifies key issues and concerns related to these conclusions that surfaced in the discussion. It should be emphasized, however, that the selection of the specific points highlighted in this section is in no way a substitute for carefully reading and weighing all of the comments made by the participants.

There was general agreement that a net impact study of school-to-work should be undertaken. However, the evaluation should not attempt to measure the impact of school-to-work as a whole but, rather, should study the impact of components of the program and the interaction among these components. In the parlance of the conferees, this approach was referred to as "disaggregating the black box."

In the discussion of "disaggregating the black box, there was considerable support for using a typology that classifies programs by their features or attributes along the lines suggested in the paper by Moffitt. An unresolved issue was how to go about selecting those features or attributes that should be studied. Approaches suggested included:

- Asking policymakers to identity program elements in which they are particularly interested.
- Using early results from the implementation study to identify and characterize school-to-work's key attributes.
- Consulting program staff and practitioners -- possibly through a focus group -- to identify interventions to be tested and outcomes anticipated.
- Looking at currently ongoing evaluations, such as the MDRC study of career academies.

Individual participants urged that specific dimensions be included in the analysis of these program components. These included intensity of treatment and the cost of individual elements of school-to-work.

There was broad support for undertaking a balanced evaluation strategy consisting of several related research components rather than conducting a single study of the school-to-work program. These components would consist of:

- An implementation study (already in progress) which, in addition to its other objectives, should be the primary vehicle (rather than the net impact study) for measuring systemic reforms resulting from school-to-work. In this connection, consideration should be given to including in the implementation study the conceptual/policy issues outlined in the paper by Dayton. Also, to assure that the study captures sufficient variation, consideration should be given to enlarging the current sample, possibly by "piggy-backing" on the fall 1996 NLSY survey.
- An observational study that would measure the net impact of individual program elements on participants.
- A small number of randomized trials to test either planned variations in the program or basic assumptions underlying the program.
- An assessment of the program's impact on employers, either through a separate survey or through one of the other components of the evaluation strategy.

There was some support for including basic research in the evaluation strategy which, through a literature review, would determine how much we already know about some of the assumptions underlying school-to-work. The review would also assess knowledge about the way the school to work transition currently operates and has operated historically in the labor market in the absence of school-to-work programs. For example, does "churning" in the labor market have a negative or positive effect on the young job seeker?

Whether through the literature review or the randomized trials, there was agreement among many participants that certain assumptions underlying school-to-work should be tested. These assumptions to be tested would include particularly: the efficacy of contextual learning, the program's effectiveness for college-bound as well as non-college bound-youth, and the possibility of increasing overall curriculum quality through school-to-work without reducing access by lower-skilled youth.

An issue of timing emerged in the debate, with some participants urging early implementation and feedback from randomized trials on basic assumptions whereas others preferred to delay these trials for several years until school-to-work matures. It was strongly urged by Moffitt and several participants that a high priority should be a baseline study as an initial step in the observational study, possibly using the fall 1996 NLSY sampling frame.

While there was general support for a quasi-experimental study to measure net impact, there was some disagreement concerning the sampling approach to be used: a nationally representative sample of youth or a sample limited to youth in school systems where school-to-work has had some time to develop. There was also some concern about identifying an adequate comparison group if, as some participants suggested, virtually all school systems have adopted or will claim to have adopted school-to-work.

There were a number of references to outcome measures in the course of the discussion. The sense of the group seemed to be that school-to-work is different from traditional training programs, such as JTPA, since school-to-work envisioned making far-reaching institutional changes in the nation's school systems. Thus, outcome measures should not be limited to the conventional measures of employment and education gains, but should also include outcomes such as whether, in response to school-to-work, employers restructure their career ladders and offer higher wage jobs to high school graduates and youth who begin to formulate concrete career plans while still in school.

Commentary on Evaluation of School-to-Work

Stephen F. Hamilton
Cornell University

The papers prepared for the roundtable discussion on evaluating the net impact of the School-to-Work Opportunities Act (STWOA), and the synopsis of the discussion, were useful and stimulating. The following commentary is conditioned by current immersion in writing a report incorporating data from a youth apprenticeship demonstration project that was directed from 1990-1995.

1. Defining the Treatment

The paper writers agree that school-to-work is not one thing. Hence, an impact evaluation cannot be useful unless it is linked closely with implementation studies so that what has had an impact is known. Just how to get inside the "treatment" is a matter on which writers differed. The direction suggested by Pouncy and elaborated by Moffitt was one of attempting to assess the impact of particular aspects or attributes of school-to-work systems. One reason to favor evaluating attributes rather than "models" is that the writers disagree with the presumption that the five models identified by Pauly et al. are distinct. It is true that representatives of these models can be found as free-standing programs, but, in a school-to-work system, youth apprenticeship might serve as one form of work-based learning for students enrolled in career academies who are studying a Tech Prep curriculum. Other students in the same school might receive their work-based learning in the form of cooperative education and still others in school-based enterprises. Moreover, all of them may have previously engaged in unpaid internships and community service to explore careers before entering apprenticeships and co-ops. The comprehensive systems envisioned by the Act cannot accurately be characterized by these five (or other) models.

The writer also has some reservations about the feasibility of identifying comparable components in programs. Work-based learning opportunities are distinguishable along five dimensions:

- Purposes-what participants are expected to learn from work
- Activities-what participants do in the workplace
- **School connections-how** closely work-based learning is related to school-based learning
- Time-how many hours are devoted to learning at work
- Cost-how expensive the program is for employers and school systems

In the following chart, the dimensions are arranged so that points to the right are more intensive and more comprehensive than points on the left. Movement from left to right is cumulative; for example, work-based learning opportunities designed to teach technical competence will also provide opportunities for career exploration and teach personal and social competence. The forms or types of work-based learning activities anchor the left and right sides of the chart, respectively. For example, visits to workplaces represent a much less intense work-based learning experience than does apprenticeship, which falls at the opposite extreme. Programs having the same form may vary along these dimensions. As another example, students engaged in a major service-learning project might spend more than 100 hours and perform complex tasks. (Hours are given for illustrative purposes, not as a strict guideline.) This portrayal of work-based learning reinforces the point that simply comparing programs featuring internships, for example, will not ensure that the internships are comparable.

Key Dimensions of Work-Based Learning Opportunities

Purposes	Exploration	Personal and social competence	Technical competence
Activities	Observation	Performing routine tasks	Planning, performing, and evaluating complex tasks
School connections	Weak, casual	Modest, occasional	Strong, multiple
Time	Short < 10 hours	Moderate 10-100 hours	Extensive > 100 hours
cost	Low	Medium	High

Similarly, it also points to difficulties in comparing almost any program attribute without examining each attribute along common criteria. For example, it would not be helpful to compare programs with and without mentoring without considering what mentoring means, how much is available, and of what quality.

2. Time Constraints

It would be difficult to maintain that Congress should not expect a report on the impact of school-to-work after four years, but anyone attempting to meet the expectation must be very careful not to attempt to deliver more than is available. The legislation explicitly calls for creating new systems that link existing institutions and build new ones, a goal that probably will not be met in any community by 1998, or even across the Nation. Some steps will have been taken and some elements put in place, but system building, by definition, is a long-term process.

Evaluators will have difficulty demonstrating in four years the kinds of impacts on young people's careers that the legislation was intended to produce. Our experience illustrates this difficulty. The first apprentices began in the demonstration as high school juniors in the fall of 1991. The plan was to prepare them for "technician-level" employment, which typically requires two years of postsecondary education. That statement means that, at this moment, five years later, only the first cohort has had a chance to complete the program as we envisioned it. Some former apprentices are now gainfully employed by their training firm in positions that are ordinarily reserved for people five to ten years older, which is defined as success. However, others are enrolled in four-year colleges and have not yet graduated. Especially for young people who were predominantly "C" students as sophomores, that program is also successful. Some have not yet completed two years of postsecondary education because they have been enrolled less than full-time or sporadically. It is acknowledged that, for some of these students, the program simply did not work. However, financial need, childbearing, illness, death in the family, and other life events have impeded the progress of others in ways that no treatment could be expected to overcome.

If an ideal system could have been designed, the program would have started even earlier so that, by grade eleven, young people would have had a stronger basis for choosing an occupational area.

If that had been accomplished, it would have been difficult to report employment impacts five years later. In addition, data on participants' work histories for five to ten years would be a much more robust test of impact than data on their first year in the full-time labor force.

Further, evaluations of the first cohorts in innovative programs may also be misleading because the programs are new. The Hawthorne effect is well known; it could have easily yielded unreproducable effects. For example, four of our apprentices were invited to the White House to witness the signing of the STWOA, a level of recognition that will not occur again. There is also danger of what might be called the pancake effect. Just as the first batch of pancakes is often inferior because the griddle was not hot enough, two or three cohorts may have to pass through a program before it runs optimally.

3. Indicators of Impact

Congress' expectation of a report after 4 years will force us to substitute proximal indicators for distal outcomes. The most obvious indicators of impact on youth are related to school performance. However, school performance as it is conventionally measured (grades, test scores, perhaps course enrollments), although important, is not adequate. Supervisors' assessments of work-related competence are a justifiable indicator of the impact of work-based learning, as are parents' and teachers' testimony and the statements of young people. However, all these data sources, though closer to the actual experience and the immediate objectives of work-based learning, suffer from being less standardized than grades and tests and form debatable real-world importance. In the project, apprentices filled in for adult employees on leave and successfully completed major tasks for their employers. The evidence of apprentices' enhanced capacity to do adult-level work was so strong that asking whether they learned at work seemed a trivial question. It would be more interesting to ask questions such as whether students learned more than they would have in class or in a different kind of program and what the cost/benefit ratio was for employers. (This assessment would have to subtract from employers' costs the value of apprentices' production because it was so substantial.) Yet, such questions are extremely difficult to answer at a reasonable cost. It is always tempting to **evaluate** what can easily be measured. The temptation to evaluate school-to-work as if it were a conventional school program should be resisted.

4. Policy Impacts

The writer supports Dayton's suggestion and would either expand it or add another category to include "institutional impacts," such as changes in the ways schools and workplaces function and the ways they interact. It was gratifying to hear from many coaches and managers that working with youth apprentices was a form of staff development. They acquired new managerial skills and a fresh perspective on their work. This result too is a benefit to employers that partly offsets their costs.

5. What Magnitude of Impact Could One Reasonably Expect From School-To-Work

Some writers referred at least obliquely to this question. It is important to look at the level of impact found in other related treatments to avoid unrealistic expectations and premature rejection. For example, what is the impact in terms of the same or similar indicators of a year of full-time, postsecondary, vocational education; of participation in the Job Corps; or of military training?

6. Meta-Analysis

The terms of implementation grants require every State to evaluate its own efforts; and when States disburse those funds to community partnerships, they surely also require evaluations. The quality of most of those evaluations will not be high, but they might usefully be mined for data and insights. Meta-analysis procedures have been refined to the point where results from multiple evaluations can be pooled to yield much greater power. Standard research reviews could identify some gems among the ore of State and local evaluations. In view of the perpetual shortage of research resources, this process could be an economical way to augment national evaluation efforts. This suggestion is related to a "quasi-experimental design" that Campbell and Stanley called "patched-up." It is argued that in cases where true experimentation was impossible, one might bring together the results of different nonexperimental studies, each compensating in some ways for the deficiencies of the other, and come out with good enough evidence.

Epilogue: Synthesis of the Papers and Discussion

I. Introduction

This chapter attempts to synthesize the commissioned papers and the roundtable discussion. This synthesis' approach is to first identify the assumptions which underlie the school-to-work legislation. These assumptions can then be stated as hypotheses for a net impact evaluation to test. We then review the methodologies which are available to test whether these assumptions are holding in school-to-work. After discussing some implementation issues, we conclude by examining the implications for school-to-work net impact evaluation.

II. The Assumptions Underlying School-to-Work

The current process for making the transition from school to work has been described as "floundering, messy, inefficient," while Asian and European systems are described as "orderly" and "smooth." The House in its report on the School-to-Work Opportunities Act, for example, noted that,

> " While our major national competitors are redefining and improving school-to-work transition systems, the United States has yet to develop one. In practical terms, this means that, unlike their peers in Japan or Germany, for example, young Americans entering the work force after high school make their way through school and into their first jobs with little guidance, direction or support. Instead of following structured career paths that provide a basis for rigorous, meaningful secondary and postsecondary education, students frequently wander aimlessly through an unchallenging, disjointed curriculum."

Thus, the overarching assumption underlying the School-to-Work Opportunities Act is that:

H-1) A structured school-to-work transition system will increase the earnings potential of those entering the workforce, smooth the transition from school-to-work and increase earnings and productivity.

However, the assumption that “orderly” school-to-work systems are better than “messy” ones has been questioned by some. Heckman (1994), for example, points out that job shopping and job search are information gathering activities. Such activities lead to matches between workers and firms that are “a major source of personal productivity enhancement with beneficial economic and social consequences.” During the roundtable, it was argued that the intent of school-to-work is for this process to take place earlier than is currently the case. However, school-to-work assumes more than just that knowledge about work should be gained before leaving school. It also assumes that this knowledge can be obtained in a school setting -- albeit altered from the present school environment -- rather than in the “real” job market. Be that as it may, there was general agreement by the roundtable that the school-to-work transition and the relative benefits of different systems are not well understood and that this is an assumption which a net impact study should seek to test.

The five program requirements listed in the Act lead to corollary assumptions about the features of a school-to-work system. The first program requirement has three parts: 1) integration of school-based and work-based learning, 2) integration of academic and vocational instruction, and 3) linking secondary and post-secondary education. The common theme which unites these elements is the elimination of distinctions among the institutions which comprise a school-to-work system.

The Senate in its report on the School-to-Work Opportunities Act stated, “The committee believes that students and businesses will benefit from a curriculum that integrates school-based and work-based learning; that is developed jointly by schools, employers and labor; and that ensures that there are high standards for graduation and that students learn the required skills.” Thus, the assumption underlying the first part of this program requirement is:

H-2) Students will learn more, retain more and be more able to apply their learning to a variety of applications if learning takes place in an integrated work-based and school-based environment.

The second part of this program requirement implies the elimination of tracking students into college-bound and vocational curricula. It is asserted that such tracking creates a stigma around vocational training, and in effect a class distinction among young people.” Thus, we have:

---

[3] See Dayton, for example.

H-3) Integrating vocational and academic curricula will eliminate the stigma often associated with vocational curricula.

The third part of this requirement is linking secondary and post-secondary education. Dayton suggests a hypothesis for this:

H-4) Linking secondary and post-secondary education will cause students to "graduate with clearly defined post-secondary goals, knowledge of educational opportunities available to pursue those goals, and perhaps a start in the next level of education."

The second program requirement is career majors. Dayton says career majors "frame the last two years of high school, providing every student with a means of focusing the curriculum around a real world post-secondary goal and learning the relationship between academic subjects and a variety of careers." It appears that the hypothesis behind career majors is:

H-5) A focus on a post-secondary goal by the eleventh grade will lead to a faster transition from secondary to post-secondary activity.

The third program requirement specifies the three components of a school-to-work system: school-based learning, work-based learning and connecting activities. Dayton points out the variety of school-based reforms implicit in school-to-work. Those not listed as separate school-to-work requirements include: articulation across grade levels, heightened awareness of technology in schools, the development and use of well-defined standards, authentic assessment and counseling, and instruction that emphasizes cooperative learning. While each of these reforms could be addressed separately, the theme of school-to-work is not the effectiveness of individual reforms, but the totality of reform of school-based learning. That is:

H-6) Reform of school-based learning will lead to students learning more and being better prepared to enter the workforce.

Pouncy and Hollister point out the different assumptions on which awork-based learning component may be designed. One of their assumptions, as expressed in the Senate report on school-to-work, is:

H-7) "[W]hen people learn concepts and skills in the process of applying them in real situations, they are more likely to retain the knowledge for use in other applications."

An alternative assumption, given by Pouncy and Hollister, is that work-based learning can provide learning that is more relevant to employers and will lead to high wage jobs. Thus:

H-8) Work-based learning develops job skills which lead to an improved transition from school-to-work and ultimately better jobs.

Connecting activities are tools for integrating work-based and school-based learning and, as such, may be more appropriate for a process evaluation than an impact evaluation. However, their inclusion as the third program requirement emphasizes the assumption that an effective school-to-work system is a partnership between employers and schools. Pouncy and Hollister point out the likely differences in results between employer-led and school-led partnerships. They hypothesize that the greater the leadership role employers have, the greater will be their incentives to contribute to a new transition system. Thus,

H-9) Employer-led school-to-work partnerships lead to more reform than school-led partnerships.

This also raises the issue of the benefits employers derive from school-to-work participation. The report accompanying the House version of the school-to-work legislation noted that,

> "Not only has the lack of school-to-work assistance had a negative impact on the earnings potential of our young people, but it also has had tremendous costs to business and our economy as a whole. Because businesses lack more highly-skilled workers, their productivity suffers and in turn, our economy as a whole suffers."

Similarly, Burtless points out that, "Employers should gain from a better trained workforce in general and from improved information about the preparation and skills of individual young people they have helped to train." Thus, we have the following hypothesis related to employer participation,

H-10) As a result of school-to-work, employers will be able to hire more skilled, more productive workers, more quickly and at less cost.

The fourth requirement is that a student be exposed to all aspects of an industry. The hypothesis underlying this requirement relates to the changing structure of work and the labor market:

H-11) Exposure to all aspects of an industry will lead to a workforce which is more adaptable to change and better prepared for the technological innovation and job instability which increasingly characterize labor markets.

The final program requirement -- equal access for all -- assumes that the system outlined in the School-to-Work Opportunities Act will benefit all students. The specific groups of students most often discussed are the college-bound, the economically disadvantaged and those with low academic skills. Foster (1995) has found that working as a teenager has a positive impact on the future earnings of non-poor teenagers but a negative impact on poor teenagers' future earnings. Even though this finding may be due to factors other than work *per se*, it raises questions about whether it can be assumed a *priori* that work experience in school, as is one aim of many school-to-work programs, will benefit both poor and non-poor students.

Pouncy and Hollister point out that the equal access requirement may conflict with the goal of an uncompromising commitment to high standards. They point out that skills differences often correlate with disadvantage, living in a high poverty area, race and other attributes. Thus, the more inclusive school-to-work is in the students it enrolls, the harder it will be to meet high standards. This leads to the assumption that school-to-work can improve the skills of low-skill youth to a degree that they can achieve the same high standards as more-skilled students.

In summary, we have the following hypotheses regarding student participation in school-to-work:

H-12) College-bound students, the economically disadvantaged and those with low academic skills will benefit from school-to-work to at least the same degree that they would benefit from education in a non-school-to-work system.

H-13) School-to-work can improve the skills of low-skill youth to a degree that they can achieve the same high standards as more-skilled students.

The assumptions identified in this section are not necessarily the only set which could be identified as underlying school-to-work. Pouncy and Hollister recommend a "limited theory-based evaluation" be included in a net impact study to identify the assumptions made by school-to-work practitioners.

However, the assumptions above serve to illustrate the range of hypotheses a net impact study might seek to test.

It is evident that the assumptions underlying school-to-work can be categorized as either system or component assumptions. System assumptions are those in which the outcomes are the result of the entire school-to-work effort. Hypotheses H-l, H-9, H-10, H-12 and H-13 fall into this category. Testing these hypotheses requires that school-to-work be fully implemented in some school districts and absent from others. Component assumptions, on the other hand, are those which isolate the affects of a single element of school-to-work. Hypotheses H-2 through H-8 and H-11 fall into this category. Testing these hypotheses requires that there be some areas where school-to-work is only partially implemented -- i.e., that there be areas where the element to be tested is missing or not available to all students. We now turn to a discussion of methodologies for testing hypotheses in each of these categories.

III. Design Issues

Net impact evaluations can be classified into three categories: 1) experimental, 2) quasi-experimental, and 3) non-experimental designs. Experiments randomly assign individuals who are eligible for a program to either a treatment or a control group. The control group is then denied access to the program being evaluated. Quasi-experiments attempt to create or find a comparison group which matches the participants in the same manner as random assignment would achieve. Non-experimental designs use. comparison groups which are acknowledged to differ from the program participants and attempt to use statistical adjustments to eliminate any bias arising from this non-comparability.

While this classification is somewhat artificial -- very seldom does an evaluation rely solely on one of these approaches -- examination of possible approaches in each category will illuminate the issues which would need to be addressed in choosing an evaluation approach for school-to-work.

Dayton describes the experimental approach as, "unarguably the strongest design." This strength comes from the small number of technical assumptions necessary to obtain unbiased estimates from an experiment. This lack of technical caveats leads to perhaps the major argument for experiments: the greater credibility results based on random assignment have with policymakers (**Manski**, 1995).

Burtless lists three criteria for randomized trials to be cost-effective:

1) " We should have reasonably good evidence suggesting that a proposed program variation can be beneficial for students or employers."

2) "The proposed treatment should hold some promise of improving outcomes in comparison with the basic school-to-work approach a school will offer."

3) "[T]here should be enough uncertainty about the benefits of the proposed alternative so that a random trial could improve the chances that the more effective option is eventually adopted."

Underlying these criteria, in part, is the consideration that using an experiment-random assignment -to evaluate school-to-workwould require denying school-to-work or one or more of its components to some students. This issue is discussed by Dayton, who agrees with opponents of random assignment that " when students become part of an experimentally designed evaluation, their futures are being determined not by ***their*** needs, but by those of the research." The counter argument to this is that if one does not know if the treatment to be tested works one is experimenting with the students in any case.

In some situations, moreover, this issue may not apply. For example, Dayton notes that, "if a program is over-enrolled, random assignment is as fair a system of deciding who gets in and who doesn't as any other." Another example is where one is comparing two treatments which are believed equally likely to be effective. In these cases, it cannot be said that the research is denying the control group the treatment believed to be most effective.

Dayton points out the major drawbacks of random assignment: it is intrusive and will be resisted by many school administrators and teachers. However, in the case of school-to-work, the introduction of random assignment may not be intrusive. The statute requires that programs "provide all students with equal access to the full range of program components." This means that, unless school-to-work is serving everyone, at some point everyone is to have an equal probability of entering the program. This is virtually the definition of random assignment.

During the roundtable, Reid Strieby described the situation of a school administrator selecting students for school-to-work participation. If the administrator chose, say, every fourth student from an alphabetical list, the administrator would be conducting an experiment. While no process in actual use is likely to be so clear cut, there may be selection processes for at least some components in a large enough number of schools that a random assignment evaluation could be conducted with little intrusion into school operations.

Related to the issue of student selection is the question of what is measured by an experiment. Heckman points out that conventional uses of random assignment in social experiments do not measure the effect of the intervention on a randomly selected person in the general population. Rather, experiments measure the impact of intention to treat. For example, if paid work experience slots are offered to 200 out of 2,000 students in the school, 50 of whom actually obtain such slots, the experiment measures the impact of offering work experience to these 200 students. While non-experimental techniques are available to measure the impacts on the 2,000 or the 50,[4] their successful application to school-to-work is open to question. The implication of this for school-to-work net impact is that random assignment may not be the most desirable approach if the "all students" assumption is to be maintained.

The experimental design for an evaluation of school-to-work systems would require randomly selecting a relatively large number of schools within a state to be school-to-work schools and randomly selecting a similar number of schools in the same state to be control schools.[5] The researcher would then track both groups of schools for several years to see if the school-to-work schools implemented programs or otherwise changed their behavior in ways different from the control schools. At the end of that period, assuming this had occurred, the researcher would begin tracking students to see if these changes led to improved outcomes. However, there are several major threats to the validity of this design which imply that it is unlikely to be feasible except in an occasional replication effort?

---

[4] See Angrist et al. (1996) for a discussion of such techniques.

[5] See Cave and Kagehiro (1995), Hollister and Hill (1995) and Raudenbush (1994) for a discussion of this approach.

1) There would likely be major spillover effects. The ideas and principles behind school-to-work are public information and well known to school administrators. Other than withholding funds from control schools, there is no way they could be prevented from implementing programs identical to school-to-work.

2) It would be impossible to prevent crossover. In order to maintain the integrity of random assignment, it would be necessary to prevent students in control schools from transferring to school-to-work schools. Since enrollment at most public schools is based on residence, parents who believed school-to-work would benefit their children could move to treatment districts, while those with negative opinions of school-to-work moved to control districts. Since school-to-work implementation is a lengthy process, this may be a major problem for an evaluation.

3) Implementation would be difficult. The design would require a state to operate dual school systems for a number of years. While many states are doing this rather than give school-to-work grants to every school district in the state, it is unlikely a state would agree to select its school-to-work grantees at random. Further, grants would have to be withheld from the control schools long enough for the treatment schools to implement school-to-work systems and graduate at least one cohort of students. This is likely to be "forever" in terms of the life span of education reforms.

It should also be noted that this design affords no opportunity for examination of individual site effects. Given that each local partnership will design a somewhat different school-to-work system, it is important to investigate which systems are more successful than others. If there is greater variation in school-to-work within states than between states, this design would not allow the evaluators to address this question.

Quasi-Experimental Designs

Quasi-experimental designs, as defined here, attempt to achieve a sample of comparison subjects whose characteristics, where they differ from those of the treatment group, are uncorrelated with

---

[6] Cave and Kagehiro (1995) report a design such as this is being used to evaluate the Comer School Project in Prince Georges County, Maryland.

school-to-work participation. There are two subcategories of quasi-experimental designs: 1) natural experiments and 2) matched comparison groups.

Natural experiments, as defined by Meyer (1995) are situations in which "there is a transparent exogenous source ofvariation in the explanatory variables that determine the treatment assignment." An example is a government policy change that allows a group of individuals to receive a service or benefit which similarly situated individuals had previously been denied. Meyer gives two criteria for such a policy change to be a natural experiment: 1) there must be a "sharp" change in policy and 2) there must not be a relationship between past values of outcomes and the policy changes.

Meyer's definition of a sharp change in policy is that the subjects of the evaluation be unable to alter their behavior in anticipation of the policy change. It is doubtful that school-to-work meets this definition. It takes years for a local school-to-work system to develop. During this development period, students at the school will experience some but not all of the effects of being in a school-to-work system. This would invalidate their role as either participants or comparisons in a natural experiment. Thus, the time span between pre- and post-school-to-work may be so long as to make it unreasonable to claim a natural experiment.

Meyer's second criterion is that there must not be a relationship between past values of outcomes and the policy changes. The achievement of this may also be doubtful in many school-to-work partnerships. For example, if a school district noticed that it was experiencing declining test scores and decided to address this problem by introducing school-to-work, the expected regression to the mean effect on test scores would violate Meyer's second criterion.

Estimating impacts in a natural experiment involves comparison either across areas or across time periods or, preferably, both. For example, the evaluator could compare the outcomes of students attending schools with school-to-work programs to outcomes of students attending schools in the same or other states without school-to-work systems. Friedlander and Robins (1995), analyzing alternative comparison groups for evaluating welfare programs found that "using individuals in one state as a comparison group for individuals in another state can lead to quite inaccurate estimates of the size of the program effect." However, they were somewhat more successful when using individuals within the same state -- either in different areas within the state or applying at different times -- as comparison groups. Thirteen percent of the cross-site pairs and 29% of the cross-cohort pairs led to different statistical inferences than the random assignment estimates. Thus, it may be worthwhile to consider comparisons within states as an approach to school-to-work net impact evaluation.

Alternatively, one could compare students who attended the school-to-work schools after school-to-work was implemented with those who attended before school-to-work implementation. In contrast to the previous design, where the source of possible bias was differences among individuals, in this design the bias is from differences in time periods. Given this difference, Meyer recommends that evaluations of natural experiments employ both designs so that neither source of variation leads to biased estimates.

The principal assumption of matched comparison groups is that, if a group matches on observable characteristics, or a function thereof, it will also match on unobservable characteristics such as motivation and innate ability. The major advantages of such an approach are that it can be implemented with less intrusion into schools' normal procedures than an experimental approach requires. As with natural experiments, one can employ both cross-sectional and pre-post observation, Dayton endorses a matched comparison group approach for school-to-work net impact. However, as he points out, "even if well done [this approach] can still leave questions about attribution of differences between participants and comparisons,"

Friedlander and Robins tound that statistical matching did not improve their estimates of program impact. They point out that only observed characteristics can be used for matching. If unobserved characteristics unrelated to those that are observed influence outcomes, matching will have little effect. Further, the same characteristics that are used in matching are usually also used as independent variables in a regression analysis to measure impacts. Thus, it may not be surprising that matching is often largely redundant.

Heckman examines five alternative matching strategies. Using data from the National JTPA Study, he concludes that simple nearest neighbor matching schemes perform very poorly when compared to estimates based on random assignment. However, he also finds that: "If . . . rich data on individual characteristics -- including labor market histories -- are available, then a nonexperimental matching estimator can be successfully implemented to estimate the impact of the program." This finding is based on samples of participants in job training and, hence, one may question its generalizability to school-to-work. Specifically, it is doubtful that labor market histories play the same role in school-to-work participation as they do in job training participation.

Non-Experimental Designs

Non-experimental approaches to estimating net impacts of interventions attempt to deal with the bias introduced by a correlation between participation in the intervention and the returns to participation by a wide variety of statistical approaches.[7] Glover and King discuss the extensive literature questioning the use of such methods. Heckman, however, points out some of the limitations of much of this research: variables were measured only on an annual basis, comparison and treatment group members did not reside in the same labor markets, insufficient information existed to determine whether comparison group members were eligible for the program, different survey instruments were used for treatments and comparisons, and no information on short-term labor force dynamics was available. Given these problems and the fact that the subjects of these studies were participants in job training and welfare programs, one may question whether the findings can be generalized to school-to-work. Nonetheless, as Burtless points out,

> "[R]esults from quasi-experimental statistical comparisons are always open to doubt. When the analysis is complete, policy makers will not know whether evaluators have obtained reliable estimates of program effectiveness for the programs included in the study. The comparison groups selected may be unconvincing, and the statistical analysis will inevitably be open to serious question."

As one non-experimental approach, Heckman discusses the widely used method of instrumental variables. This method requires finding a variable which is correlated with participation but is not otherwise a determinant of the outcome of interest. Heckman shows that this assumes that "persons do not make their decisions to participate in the program based on unobserved or forecastable components of program gains." In the school-to-work context, this condition may be satisfied if school-to-work students are selected by school administrators. However, it is perhaps more likely that -- if available "slots" are scarce -- administrators will choose the students they deem most likely to benefit. Such choice may violate Heckman's condition. Be that as it may, Heckman points out that the issue of the necessary behavioral assumptions underlying the instrumental variables method "cannot be settled by a statistical analysis." To use instrumental variables, the researcher must have a thorough understanding of the participation decision.

---

[7] See Heckman and Robb (1986) for a discussion of such methods.

Moffitt proposes an innovative non-experimental approach to school-to-work net impact evaluation based on a factorial design. His approach would involve defining an attribute-based typology of school-to-work programs, placing partnerships within this typology and then exploiting the natural variation among partnerships to determine the net impact of an attribute holding other attributes constant. An advantage of such an approach is that scale and intensity can be defined as attributes and evaluated. The major drawback to Moffitt's approach is the potentially large number of attributes which may differ among partnerships. This could lead to an unmanageably complex design. If that issue is surmountable, however, this strategy, coupled with randomized trials as Moffitt suggests, would "yield a comprehensive net impact analysis incorporating all important sets of influences and factors."

IV. Implementation

Nearly all the roundtable participants endorsed the idea of developing a typology of school-to-work systems as a first step towards a net impact evaluation of school-to-work. However, review of the typologies which have been discussed to date suggests this will not be an easy task. King and Glover discuss and expand the five categories of models of school-to-work defined by Pauly et al. (1994). They conclude that there are two "typologies" of school-to-work programs: school reform and occupational preparation. They further identify the "key indicators distinguishing these two typologies" as the roles of work-based learning and preparation in job specific skills.

It should be noted, however, that others have criticized the classification of Pauly et al. (1994). School-to-work is intended to encompass both occupational preparation and education reform. Thus, Hamilton argues that the Pauly et al. (1994) "models" are themselves components of a school-to-work program rather than models of a school-to-work system. Similarly, Pouncy and Hollister provide a basis for rating the degree to which such models include the essential components of school-to-work. Of the six "main types of school-to-work programs" they rate, youth apprenticeship scores seven out of a possible nine points while the others score only three to five points. Examination of the categories suggests that components defined in this way focus on only one aspect of school-to-work: work-based learning.

An alternative approach to selecting components for evaluation would focus on "critical elements." Pauly et al. (1994) and Stern et al. (1995) provide possible lists. While the lists don't entirely agree, most of the elements on each list are concerned with aspects of integration and linkages -- of academic and occupational learning and of school- and work-based learning, for example. Although integration and linkages require that several components be present to integrate or link, this categorization of school-to-work systems may also lead to an evaluation focussed on one aspect of school-to-work.

Another key implementation issue is the timing of the evaluation. Everyone seems to agree that fully implementing a school-to-work system is a long process, requiring at least five and possibly ten years. If one believes that school-to-work began when the first grants were awarded in 1994, this means evaluation should begin no earlier than 1999. Given the time necessary for a student to complete a school-to-work program and make the transition into employment, net impact results would not be available until about 2005. Glover and King point out that,

> "The combination of long program and post-program time frames means that the eventual findings lose some of their currency and relevance. It is doubtful whether evaluation results retain their timeliness after a decade or more. Their policy shelf-life is likely to be far shorter."

Rebecca Maynard, however, questioned whether net impact evaluation should wait for full implementation. She pointed out that the longer it takes to get results, the more schools and society will have invested in school-to-work. This will create institutional inertia making it difficult to apply the evaluation results.

In any case, even if the decision is that more time must be allowed for school-to-work systems to develop before a net impact evaluation can begin, this does not mean that no work should be done on a net impact evaluation before then. Several authors discuss activities which should be undertaken before a net impact evaluation begins. King and Glover recommend "detailed process evaluations before sites are selected or particular school-to-work interventions/models are selected." While much of the information needed for such studies will be collected by the other evaluation activities described in the introduction, there will still be a necessity to supplement this information and analyze it from a net impact evaluation perspective. Dayton points out that "an important element of examining participant impact is looking at pre-post changes." He further notes that much of this information may be lost if not obtained "early on." Even if "pre" data can be collected retrospectively, it would be necessary to determine that the data exist before committing to an evaluation design.

With regard to more mundane -- but no less important -- matters, **Heckman** identifies four sources of bias in net impact estimates:

1) drawing participants and comparisons from different labor markets,
2) administering different questionnaires to the two groups,
3) differences in unobserved characteristics between the two groups (selection bias),
4) differences in observed characteristics.

With the exception of selection bias, these sources of bias are to a large extent within the control of evaluation implementors. While selection bias receives the overwhelmingly major portion of attention in the literature, Heckman's empirical work on JTPA underscores the importance of paying careful attention to these other issues in implementing an evaluation.

V. Conclusions

There was general agreement among the roundtable participants that a net impact evaluation of school-to-work should be conducted. The most compelling case for such an evaluation was mentioned by Rebecca Maynard: the human capital costs of implementing major reforms without the evidence net impact evaluation can provide. A prominent example is California's experience with the "whole language" method of teaching reading. After this method was implemented, California's reading scores plummeted. Yet it took a decade to reverse the decision and return to the phonics approach (New *York* Times, 1996).

As noted above, the issues which an evaluation might address divide into two categories: systemic issues and component issues. There was broad support among the roundtable for a balanced evaluation strategy addressing both classes of issues. However, if this is not feasible within a reasonable evaluation budget, there was no agreement on which set of issues should receive priority.

If either or both categories of issues are to be addressed, the first step in an evaluation would be to define what is to be evaluated. In the systems evaluation, this requires answering the two questions: 1) What is school-to-work? and 2) To what should it be compared? The first question may, in practice, be a timing question: When has a state's school-to-work system developed sufficiently to call it school-to-work? The usual answer to the second question is: What the school system would be if it had no school-to-work implementation grant. However this may be very hard to determine. Rebecca Maynard pointed out during the roundtable that schools are always being "reformed." If school-to-work wasn't being implemented, some other education reform likely would be.

For a component evaluation , the same two questions must be answered except "school-to-work" in question 1 is replaced by "the component to be evaluated." Hamilton argues that to be helpful, the answer to this question should consider not just the definition of the component but also the quantity and quality. The second question defines the margin at which the component is evaluated. The two extremes are: the component added to a system which has none of the elements of school-to-work and comparison of a school system with all the elements of school-to-work compared to a system missing only the component to be evaluated. The approach to answering these questions discussed in the roundtable is to develop a typology of school-to-work systems and programs. This will not be a simple task. To serve for evaluation design purposes, a typology would need to capture all elements of a school-to-work system and the significant dimensions along which each element varies. No previously developed school-to-work typology meets this criterion.

The reviews and critiques in the papers of various approaches to evaluation design identify important benefits and caveats for each approach considered. Burtless points out that, "The potential benefits of an experiment are clearest when the focus of study is narrow." This, and the discussion above, suggest that random assignment should be used if an evaluation of components -- especially an evaluation of components that are unlikely to have enough slots to accommodate every student --

is conducted, but an evaluation addressing system issues should be based on a quasi-experimental design.

Employers warrant special attention in school-to-work evaluation. Not only do they play an important role in school-to-work, but net impact evidence is needed to persuade employers that participation in school-to-work makes good business sense. As Dayton says,

> "[E]mployers in the United States have generally not regarded it as part of their responsibility to help prepare young people for work...With the emphasis in the STWOA on work-based learning, this raises an important question concerning the degree to which employers will be willing to contribute the time and energy needed to develop and support such a system, which will benefit the general good but perhaps not serve their particular needs in any immediate fashion. With the emphasis on short-term profits, competition and the trend toward downsizing in this country in the last decade, this issue is particularly relevant."

Finally, it is an open question as to when a net impact evaluation should begin. The implementation evaluation currently underway will provide some information on the development of school-to-work systems. This may be sufficient to answer the question of when a systems evaluation should begin. However, a components evaluation can and should begin before school-to-work is fully implemented. Thus, the prior question is whether net impact evaluation of school-to-work is to be limited to system-wide issues. Whatever the answer to this question, however, if a net impact evaluation of school-to-work is to be conducted, work on its design should begin immediately.

Bibliography

Academy for Educational Development. 1996. ***School to Work Making the Transition.*** Washington, D.C.: Academy for Educational Development.

Anderson, Elijah. 1990. "Racial Tension, Cultural Conflicts, and Problems of Employment Training Programs." ***The Nature of Work.*** Kai Erikson and Steven Peter Vallas, editors. New Haven, Connecticut: Yale University Press.

Angrist, Joshua. 1990. "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence From Social Security Administration Records," ***American Economic Review, 80:*** 313-335.

Angrist, Joshua, and Guido Imbens. 1991. "Sources of Identifying Information in Selection Models," NBER Technical Working Paper 117.

Angrist, Joshua, Guido Imbens, and Donald Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables," ***Journal of the American Statistical Association,*** 91(434): 444-455.

Ashenfelter, Orley. 1978. "Estimating the Effect of Training Programs on Earnings," ***Review of Economics and Statistics,*** 40(1): 47-57.

Ashenfelter, O., and D. Card. 1985. "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs," ***Review of Economics and Statistics, 67: 648-660.***

Bailey, Thomas R., ed. 1995. ***Learning to Work Employer Involvement in School-To-Work Transition Programs.*** Washington, D. C.: The Brookings Institution.

Bailey, Thomas R., and Donna Merritt. ***1993. The School-To-Work Transition and Youth Apprenticeship: Lessons from the U.S. Experience.*** New York: Manpower Demonstration Research Corporation.

Bailis, Lawrence Neil. 1995. ***Evaluation of Walks of Life: Second Annual Report.*** Waltham, Massachusetts: Brandeis University.

Barnow, Burt S. 1987. "The Impact of CETA Programs on Earnings: A Review of the Literature." ***Journal of Human Resources,*** 22(2): 157-193.

Barton, Paul E. ***1994. Indicators of the School-to-Work Transition.*** Princeton, New Jersey: Policy Information Center, Educational Testing Service.

Bell, Stephen H., Larry L. Orr, John D. Blomquist, and Glen D. Cain. 1995. *Program Applicants as a Comparison Group in Evaluating Training Programs: Theory and a Test.* Kalamazoo, Michigan: W. E. Upjohn Institute for Employment Research.

Berryman, Susan, and Thomas Bailey. 1992. *The Double Helix of Education and the Economy. New* York: Institute on Education and the Economy, Teachers College, Columbia University.

Betsey, Charles L., Robinson G. Hollister, Jr., and Mary R. Papageorgiou (eds.). 1985. Youth *Employment and Training Programs: The YEDPA Years.* Washington, D. C.: National Academy Press.

Bishop, John. 1992. "High School Performance and Employee Recruitment." *Journal of Labor Research,* 13 (Winter): 41-4.

Bjorklund, Anders, and Robert Moffitt. 1987. "Estimation of Wage Gains and Welfare Gains in Self-Selection Models," *Review of Economics and Statistics,* 49(1): 42-49.

Blalock, Ann Bonar (ed.). 1990. *Evaluating Social Programs at the State and Local Level: The JTPA Evaluation Design Project.* Kalamazoo, Michigan: W. E. Upjohn Institute for Employment Research.

Blanchflower, D., and A. Oswald. 1994. The *Wage Curve,* Cambridge, Massachusetts: Massachusetts Institute of Technology Press.

Bloom, Howard. 1984. "Accounting For No-Shows in Experimental Evaluation Designs," *Evaluation Review, 82 (2): 225-246.*

Bloom, Howard S., Larry L. Orr, George Cave, Stephen H. Bell and Fred Doolittle. 1993. *The National JTPA Study: Title IIA Impacts on Employment and Earnings at 18 Months.* Washington, D. C.: U. S. Department of Labor, Employment and Training Administration.

Boesel, David, and Laurel McFarland. 1994. *National Assessment of Vocational Education, Final Report to Congress.* Washington, D. C.: U. S. Department of Education, Office of Research.

Bourgois, Philippe I. 1995. *In Search of Respect: Selling Crack in El Barrio.* Cambridge; New York: Cambridge University Press.

Bowman, William R. 1992. *Evaluating JTPA Programs for Economically Disadvantaged Adults: A Case Study of Utah and General Findings.* Washington, D. C.: National Commission for Employment Policy Research Report 92-02.

Bragg, Debra, and R.E. Hamm. 1995. *Linking College and Work Exemplary Practices in Two-Year College Work-Based Learning Programs.* Berkeley, California: National Center for Research in Vocational Education, University of California.

Bragg, Debra, James D. Layton, and F. T. Hammons. 1994. *Tech Prep Implementation in the United States: Promising Trends and Lingering Challenges.* Berkeley, California: National Center for Research on Vocational Education, University of California.

Brustein, Michael, and Marty Mahler. 1994. ***AVA Guide to the School-to-Work Opportunities Act.*** Alexandria, Virginia: American Vocational Association.

Bryant, Edward and Kalman Rupp. 1986. "Evaluating the Impact of CETA on Participant Earnings," *Evaluation Review,* 11(4): 473-492.

Burtless, Gary. 1995. "The Case for Randomized Field Trials in Economic and Policy Research," *Journal of Economic Perspectives,* 9(2): 63-84.

California Business Roundtable. 1994. *Mobilizing for Competitiveness: Linking Education and Training to Jobs.*

California Department of Education. 1994-95. *Fact Book.* Sacramento, California.

California Department of Education. 1994-95. *Career Path Guides in Agriculture, Business, Health Careers, Consumer/ Home Economics, and Industrial & Technology.* Sacramento, California.

California Department of Education. 1992. *Second to None: A Vision of the New California High School. The Report of the California High School Task Force.* Sacramento, California.

California Postsecondary Education Commission. 1994. *The Performance of California Higher Education.* Sacramento, California.

Cave, George and Susie Kagehiro. 1995. *Accelerated Middle Schools: Assessing the Feasibility of a Net Impact Evaluation,* New York, New York: Manpower Demonstration Research Corporation.

Clements, Nancy, James Heckman, and Jeffrey Smith. 1997. "Making the Most Out of Social Experiments: Reducing the Intrinsic Uncertainty in Evidence from Randomized Trials with an Application to the National JTPA Experiment," *Review of Economic Studies,* forthcoming.

Cochrane, William G., and Donald Rubin. 1973. "Controlling Bias in Observational Studies," *Sankyha,* 35: 417-446.

Commission on Skills of the American Workforce. 1990. *America's Choice: High Skills or Low Wages.* Washington, D. C.: National Center on Education and the Economy.

Commission on Work, Family, and Citizenship. 1988. *The Forgotten Half: Non-College Youth in America* and *The Forgotten Half: Pathways to Success for America's Youth and Young Families.* Washington, D. C.: William T. Grant Foundation.

Couch, Kenneth. 1992. "New Evidence on the Long-term Effects of Employment Training Programs." *Journal of Labor Economics,* 10(4): 380-388.

Council of Chief State School Officers. 1995. *Building School-to-Work Transition Systems in Eight States. Final Report.* Washington. D. C.: Council of Chief State School Officers.

Cox, David. 1958. *The Planning of Experiments. New* York, New York: Wiley.

Crain, R.L., A.L. Heebner, and Y-P Si. 1992. *The Effectiveness of New York City's Career Magnet Schools: An Evaluation of Ninth Grade Performance Using an Experimental Design.* National Center for Research and Vocational Education. Publication No. MDS-173.

Daggett, Willard R. 1993. "Answering the Call for School Reform." *The Balance Sheet.*

Dayton, C. 1995. *California Partnership Academies: 1993-94 Evaluation Report.* Sacramento, California: California Department of Education.

Dayton, C. and M. Rahn. 1994. *A Report on the California New Youth Apprenticeship Project.* Nevada City, California: Foothill Associates.

Dayton, C., et al. 1992. "The California Partnership Academies: Remembering the Forgotten Half." *Phi Delta Kappan,* March, 539-545.

Dayton, C. A. Weisberg, and D. Stern. *1989. California Partnership Academies: 1987-88 Evaluation Report.* Berkeley, California: Policy Analysis for California Education (PACE).

Devine, T., and J. Heckman. 1994. "Consequences of Eligibility Rules for a Social Program: A Study of the Job Training Partnership Act," forthcoming in *Research in Labor Economics,* edited by S. Polachek. Bridgeport, Connecticut: JAI.

Dickenson, K., T. Johnson, and R. West. 1986. "An Analysis of the Sensitivity of Quasi-Experimental Net Impact Estimates of CETA Programs," *Evaluation Review,* 11(4): 452-472.

Doolittle, Fred, and Linda Traeger. 1990. *Implementing the National JTPA Study. New* York, New York: Manpower Demonstration Research Corporation.

Ellickson, Phyllis, and Robert Bell. 1990. "Drug Prevention in Junior High: A Multi-Site Longitudinal Test." *Science, 247,* pp. 1299-1305.

Fan, Jianqing. 1993. "Local Linear Regression Smoothers and Their Minimax Efficiencies," *The Annals of Statistics,* 21 (1): 196-216.

Foster, E. Michael. 1995. "Why Teens do not Benefit from Work Experience Programs: Evidence from Brother Comparisons," *Journal of Public Analysis and Management,* 14(3):393-414.

Fraker, Thomas, and Rebecca Maynard. 1987. "The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs," *Journal of Human Resources,* 22:194-227.

Friedlander, Daniel, and Gary Burtless. 1995. *Five Years After: The Long-Term Effects of Welfare-to-Work Programs.* New York, New York: Russell Sage Foundation.

Fullan, M. 1995. "Managing Change." California High School Networks Project. *Network News,* June 15.

Garfinkel, Irwin, Charles F. Manski, and Charles Michalopoulos. 1991. "Are Micro-Experiments Always Best? Randomization of Individuals or Sites." In Charles Manski and Irvwin Garfinkel (eds.), ***Evaluating Welfare and Training Programs.*** Cambridge, Massachusetts: Harvard University Press.

Goldberger, Susan, and Richard Kazis. 1995. ***Revitalizing High Schools: What the School-to-Career Movement Can Contribute.*** Washington, D. C.: American Youth Policy Forum, et al.

Goldberger, Susan, Richard Kazis, and Mary Kathleen O'Flanagan. ***1994. Learning Through Work Designing and Implementing Quality Worksite Learning for High School Students.*** *New* York, New York: Manpower Demonstration Research Corporation.

Goldberger, Susan. 1992. ***From High School to High-Skilled Health Careers: New Models of Work-and-learning in Health Cure.*** Boston, Massachusetts: Jobs for the Future.

Goldberger, Susan. 1993. ***Creating an American-Style Youth Apprenticeship Program: A Formative Evaluation of Project ProTech.*** Boston, Massachusetts: Jobs for the Future.

Goldfeld, Steven, and Richard Quandt. 1972. ***Nonlinear Methods in Econometrics,*** Amsterdam: North Holland.

William T. Grant Foundation. 1988. ***The Forgotten Half Noncollege Youth in America: An Interim Report on the School-to-Work Transition.*** Washington, D. C.: Youth and America's Future, the William T. Grant Foundation's Commission on Work, Family, and Citizenship.

Grubb, W. Norton, and Norena Badway. 1995. "Linking School-Based and Work-Based Learning: The Implications of LaGuardia's Co-op Seminars for School-to-Work Programs." Paper prepared for the Office of Technology Assessment, U.S. Congress.

Grubb, W. Norton (ed.) 1995. ***Education Through Occupations in American High Schools. New*** York, New York: Teachers College Press.

Grubb, W. Norton, Gary Davis, Jeannie Lum, Jane Plihal and Carol Morgraine. 1991. ***The Cunning Hand, The Cultured Mind : Models for Integrated Vocational and Academic Education.*** Berkeley, California: National Center for Research in Vocational Education.

Gruber, David. 1992. ***Toward a Seamless System for Youth Development: A New Strategy for Integrating Resources, Programs, and Institutions.*** Boston, Massachusetts: Jobs for the Future.

Hamilton, Mary Agnes, and Stephen F. Hamilton. ***1994. Increasing Adolescents' Planful Competence:A Report on the Cornell Youth Apprenticeship Demonstration Project.*** Ithaca, New York: Cornell University Youth and Work Program.

Hamilton, Mary Agnes, and Stephen F. Hamilton. 1993. ***Toward a Youth Apprenticeship System.*** Ithaca, New York: Department of Human Development and Family Studies, Cornell University.

Hamilton, Stephen F., Mary Agnes Hamilton, and Benjamin J. Wood. 1991. ***Creating Apprenticeship Opportunities for Youth.*** Ithaca, New York: Department of Human Development and Family Studies, Cornell University.

Hamilton, Stephen F., and Mary Agnes Hamilton. 1994. ***Opening Career Paths for Youth: What Needs to** be **Done? Who can do it?*** Washington, D. C.: Cornell University Youth and Work Program.

Härdle, Wolfgang. 1990. ***Applied Nonparametric Regression,*** Cambridge, England: Cambridge University Press.

Hayward, B.J. et al. 1992. ***Evaluation of Dropout Prevention and Reentry Demonstration Projects in Vocational Education. Final Report: Phase II.*** Research Triangle, North Carolina: Research Triangle Institute. [draft]

Heckman, James. 1978. "Dummy Endogenous Variables in a Simultaneous Equations System," ***Econometrica,*** 46(4): 931-959.

Heckman, James. 1992. "Randomization and Social Program," In ***Evaluating Welfare and Training Programs,*** edited by C. Manski and I. Garfinkle. Cambridge, Massachusetts: Harvard University Press.

Heckman, James. 1995. "Instrumental Variables: A Study of Implicit Behavioral Assumptions in One Widely-Used Estimator," unpublished manuscript, University of Chicago.

Heckman, James, and Bo Honoré. 1990. "Empirical Content of the Roy Model," ***Econometrica,*** 58: 1121-1149.

Heckman, James J., and V. Joseph Hotz. 1989. "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training." ***Journal of the American Statistical Association,*** 84(408): 862-874.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. 1995a. "Nonparametric Characterization of Selection Bias Using Experimental Data, Part I: Definitions, Applications and Empirical Results," unpublished manuscript, University of Chicago.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. 1995b. "Nonparametric Characterization of Selection Bias Using Experimental Data, Part II: Econometric Theory and Methods and Monte Carlo Evidence," ***Econometrica,*** under revision.

Heckman, James, Hidehiko Ichimura, and Petra Todd. 1997a. "Matching as an Econometric Evaluation Estimator: Theory and Evidence on Its Performance Applied to the JTPA Program, Part I: Theory and Methods," ***Review of Economics Studies,*** forthcoming.

Heckman, James, Hidehiko Ichimura, and Petra Todd. 1997b. "Matching as an Econometric Evaluation Estimator: Theory and Evidence Applied to the JTPA Program, Part II: Empirical Evidence," *Review of Economics Studies*, forthcoming.

Heckman, James, Hidehiko Ichimura, and Petra Todd. 1996. "Sources if Selection bias in Evaluating Social Programs: An Interpretation of Conventional Measures and Evidence of the Effectiveness of Matching as a Program Evaluation Method." *Proceedings of the National Academy of Sciences*, 93:13146-13150.

Heckman, James, and Richard Robb. 1985, "Alternative Methods for Evaluating the Impact of Interventions," In *Longitudinal Analysis of Labor Market Data, New* York, New York: Wiley.

Heckman, James, and Richard Robb. 1986. "Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes," In *Drawing Inferences From Self-Selected Samples,* edited by H. Wainer. Berlin, Germany: Springer Verlag.

Heckman, James, and Rebecca Roselius. 1994. "Evaluating the Impact of Training on the Earnings and Labor Force Status of Women: Better Data Help a Lot," unpublished manuscript, University of Chicago.

Heckman, James, and Jeffrey Smith. 1993. "Assessing the Case for Randomized Evaluations of Social Programs," In *Measuring Labour Market Outcomes,* edited by K. Jensen and P. K. Madsen. Copenhagen, Denmark: Ministry of Labor.

Heckman, James J., and Jeffrey A. Smith. 1995. "Assessing the Case for Social Experiments," *Journal of Economic Perspectives,* 9(2): 85-110.

Heckman, James, and Jeffrey Smith. 1994. "Determinants of Participation in a Social Program: Evidence from JTPA," unpublished manuscript, University of Chicago.

Heckman, James, Jeffrey Smith, and Christopher Taber. 1997. "Accounting For Dropouts in Evaluations of Social Experiments," *Review of Economics and Statistics,* forthcoming.

Hershey, Alan, Marsha Silverberg, and Tom Owens. *1995. The Diverse Forms of Tech-Prep: Implementation Approaches in Ten Local Consortia.* Princeton, New Jersey: Mathematica Policy Research, Inc.

Hoachlander, E.G. *1994. Industry-Based Education: A New Approach for School-to- WorkTransition*. Berkeley, California: MPR Associates.

Hoerner, J.L., et al. 1992. *Tech Prep: An Embryonic Idea and Divergent Practice.* Berkeley, California: National Center for Research in Vocational Education.

Hoffinger, Alex, and Charles Goldberg. *1995. Connecting Activities in School-to-Career Programs: A User's Manual.* Boston, Massachusetts: Bay State Skills Corporation.

Hollister, Robinson, and Jennifer Hill. 1995. "Problems in the Evaluation of Community-Wide Initiatives." In *New Approaches to Evaluating Community Initiatives: Concepts, Methods, and*

BEST COPY AVAILABLE

Contexts," edited by James P. Connell, Anne C. Kubisch, Lisbeth Schorr, and Carol Weiss. Queenstown, Maryland: The Aspen Institute.

Hotz, V. Joseph. 1992. "Designing an Evaluation of the Job Training Partnership Act." In *Evaluating Welfare and Training Programs,* edited by C. Manski and I. Garfinkel. Cambridge, Massachusetts: Harvard University Press.

Hotz, V. Joseph, and Seth Sanders. *1994. Bounding Treatment Effects in Controlled and Natural Experiments Subject to Post-Randomization Treatment Choice,* Population Research Center, University of Chicago.

Ihlanfeldt, Keith R., and David L. Sjoquist. 1990. "Job Accessibility and Racial Differences in Youth Employment Rates." *The American Economic Review, 80: 267-76.*

Ihlanfeldt, Keith R., and David L. Sjoquist. 1991. 'The Effect of Job Access on Black and White Youth Employment: A Cross-Sectional Analysis." *Urban Studies, 28: 255-65.*

Imbens, Guido, and Joshua Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects," *Econometrica, 62 (4): 467-476.*

Jackson, Russell H., Micah Dial, and Rhonda Strauss. *1994. Evaluation of Tech Prep System Development and Implementation in Texas Public Schools and Institutions of Higher Education.* Final Report 1993-94. Houston, Texas: Decision Information Resources, Inc.

Jacobs, Heidi Hayes. *1989. Interdisciplinary Curriculum: Design and Implementation.* Alexandria, Virginia: ASCD.

Jacobs, Lauren. *1995. The School-to-Work Opportunities Act of 1994: A Guide to the Law and How to Use It.* Washington, D. C.: Center for Law and Education.

Judge, George, William Griffiths, R. Carter Hill, and Tsoung-Chao *Lee.* 1980. *The Theory and Practice of Econometrics,* New York, New York: John Wiley.

Kazis, Richard. *1993. Improving the Transition from School to Work in the United States.* Washington, D. C.: Jobs for the Future.

Kemple, James J. and JoAnn Leah Rock. *1996. Career Academies: Early Lessons from a IO-Site Evaluation. New* York, New York: Manpower Demonstration Research Corporatopn.

Kemple, James J., Fred Dolittle, and John Wallace. *1993. The National JTPA Study: Site Characteristics in Participation Patterns,* New York, New York: Manpower Demonstration Research Corporation.

King, Christopher T., and Leslie 0. Lawson. 1996. "Measuring Entry-Level Employment by Occupation: A Review and Recommendations." Austin: Center for the Study of Human Resources, The University of Texas.

Kopp, Hilary, and Richard Kazis, with Andrew Churchill. 1995. ***Promising Practices: A Study of Ten School-to-Career Programs.*** Boston, Massachusetts: Jobs for the Future.

LaLonde. Robert. June 1986. "Evaluating The Econometric Evaluation of Training Programs," ***American Economic Review, 76: 604-620.***

LaLonde, Robert J. 1995. "The Promise of Public Sector-Sponsored Training Programs." ***Journal of Economic Perspectives,*** 9(2).

Licht, Walter. 1992. ***Getting Work: Philadelphia, 1840-1950.*** Cambridge, Massachusetts: Harvard University Press.

Lerman, Robert I., and Hillard Pouncy. 1990. "The Compelling Case for Youth Apprenticeships." *The Public Interest, 101* (Fall): 62-77.

MacAllum, Keith, and Patricia Ma. ***1995. Skills, Standards and Entry Level Work Elements of a Strategy for Youth Employability Development.*** Washington, D. C.: U.S. Department of Labor, Employment and Training Administration.

Majors, Richard, and Janet Mancini-Billson. ***1992. Cool Pose: The Dilemmas of Black Manhood in America.*** New York, New York: Lexington Books.

Manpower Demonstration Research Corporation, Abt Associates, New York University, National Opinion Research Corporation and ICF, Inc. 1990. ***Design of the National JTPA Study.*** Bethesda, Maryland: Abt Associates.

Manski, Charles F. 1995. "Learning About Social Programs from Experiments with Random Assignment of Treatments," Institute for Research on Poverty, Discussion Paper Number ***1061-95.***

Manski, C., and S. Lerman. 1977. "The Estimation of Choice Probabilities from Choice-Based Samples," ***Econometrica, 45: 1977-1988.***

Mathematica Policy Research, Inc. ***1994. The School-to-Work/Youth Apprenticeship Demonstration: Preliminary Findings.*** Research and Evaluation Report Series 94-E. Washington, D. C.: U.S. Department of Labor, Employment and Training Administration.

Mathematica Policy Research, Inc. 1995. "Evaluation of School-to-Work Implementation Project Description." Mimeograph.

Moffitt, R. 1991. "Program Evaluation with Nonexperimental Data." ***Evaluation Review,*** 15: ***291-314.***

Moffitt, R. 1992. "Evaluation Methods for Program Entry Effects." In ***Evaluating Welfare and Training Programs,*** edited by C. Manski and I. Garfinkel. Cambridge, Massachusetts: Harvard University Press.

Moffitt, R. 1995. "Selection-Bias Adjustment in Treatment-Effect Models as a Method of Aggregation," ***1995 Proceedings of the American Statistical Association,*** forthcoming.

National Association of Secondary School Principals. 1996. ***Breaking Ranks.*** Princeton, New Jersey: Carnegie Foundation for the Advancement of Teaching and the National Association of Secondary School Principals.

National Center for Research in Vocational Education. 1994. ***School-to-Work Facts.*** Berkeley, California.

National Center for, Research in Vocational Education. 1995. ***Getting to Work A Guide for Better Schools.*** Berkeley, California.

National Center for Research in Vocational Education. 1995. ***Legislative Principles for Career-Related Education and Training: What Research Supports.*** Berkeley, California: National Center for Research in Vocational Education, University of California.

National Center on Education and the Economy. ***1990. America's Choice: High Skills or Low Wages!*** The Report of the Commission on the Skills of the American Workforce. Rochester, New York.

National Center on Education and the Economy and the Learning Research and Development Center. ***1996. New Standards: Performance Standards for English Language Arts, Mathematics, Science and Applied Learning.*** Washington, D. C.:National Center on Education and the Economy.

National Commission for Employment Policy. 1991. ***A Feasibility Study of the Use of Unemployment Insurance Wage-Record Data as an Evaluation Tool for JTPA: Report on Project's Phase I Activities.*** Washington, D. C.: National Commission for Employment Policy, Research Report 90-02.

National Commission for Employment Policy. 1992. ***Using Unemployment Insurance Wage-Record Data for JTPA Performance Management.*** Washington, D. C.: National Commission for Employment Policy Research Report 91-07.

National Governors Association. 1995. "State Progress in School-to-Work System Development." ***StateLine*** (July 29). Washington, D. C.: National Governors Association.

The New York Times. May 22, 1996. "California Leads Revival of Teaching by Phonics."

Oakes, J. ***1985. Keeping Track How Schools Structure Inequality.*** New Haven, Connecticut: Yale University Press.

O'Neill, John. 1995. "On Lasting School Reform: A Conversation with Ted Sizer." ***Educational Leadership.***

Orr, Margaret Terry. *1996. Wisconsin Youth Apprenticeship Program in Printing: Evaluation 1993-1995.* Boston. Massachusetts: Jobs for the Future.

Owens, Thomas R. et al. *1995. Washington State School-to-Work Evaluation.* 3 vols. Portland, Oregon: Northwest Regional Educational Laboratory.

Parnell, Dale. *1985. The Neglected Majority.* Washington, D. C.: The Community College Press.

Parnell, Dale. 1993. "What is the Tech Prep/Associate Degree Program?" *The Balance Sheet,* Winter 1993.

Pauly, Edward, and Deborah E. Thompson. *1993. Assisting Schools and Disadvantaged Children by Getting and Using Better Evidence on What Works in Chapter 1: The Opportunities and Limitations of Field Tests Using Random Assignment.* New York, New York: Manpower Demonstration Research Corporation.

Pauly, Edward, Hilary Kopp, and Joshua Haimson. *1994. Home-Grown Lessons: Innovative Programs Linking Work and High School.* New York, New York: Manpower Demonstration Research Corporation.

Policy Analysis for California Education. 1992-93. *Conditions of Education in California.* Berkeley, California.

Policy Analysis for California Education. *1995. Conditions of Education in California, 1994-95.* Berkeley, California.

Pouncy, Hillard, and Andrew Hahn. 1996. "The Urban Poverty/School-to-Career Transition Connection: A Reconnaissance To Explore Implications for Grant-Making Strategy." Mimeograph.

Pouncy, Hillard, and Ronald B Mincy. 1994. "Out of Welfare Strategies for Welfare-Bound Youth." In *The Work Alternative,* edited by Demetra Smith Nightingale and Robert H. Haveman. Washington, D. C.: The Urban Institute Press.

Quandt, Richard. 1972. "A New Approach to Estimating Switching Regressions," *Journal of the American Statistical Association,* 67:306-310.

Raizen, S.A. *1989. Reforming Education for Work: A Cognitive Science Perspective.* Berkeley, California: National Center for Research in Vocational Education.

Rao, C. 1965. "On Discrete Distributions Arising Out of Methods of Ascertainment," In *Classical and Contagious Discrete Distributions,* edited by G. Patil. New York, New York: Pergamon Press.

Raudenbush, Stephen W. 1994. *Statistical Analysis and Optimal Design in Cluster Randomized Trials.*

Reisner, Elizabeth R., Nancy E. Adelman, John S. Breckinridge and Christine D. Kulick. 1995. ***Early Progress in States with Implementation Grants Under the School-to-Work Opportunities Act,*** Washington, D. C.: Policy Studies Associates.

Resnick, L.B. 1987. "Learning in School and Out." ***Educational Researcher,*** 16:13-20.

Rivlin, Alice, and Michael Timpane. 1975. "Planned Variation in Education: An Assessment" in ***Planned Variation in Education: Should We Give up or Try Harder?*** Alice Rivlin and Michael Timpane, eds. Washington, D. C.: Brookings Institution.

Roselius, Rebecca. 1995. "Evaluating Social Program Evaluation Methods: New Evidence on What Works, What Doesn't, and Why," unpublished manuscript, University of Chicago.

Rosenbaum, Paul, and Donald Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies For Causal Effects," ***Biometrika, 70: 41-55.***

Rosenbaum, Paul, and Donald Rubin. 1985. "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score," ***American Statistician, 39(1): 33-38.***

Rossi, P., and H. Freeman. ***1993. Evaluation: A Systematic Approach.*** Newbury Park, California: Sage Publications.

Roy, Andrew D. 1951. "Some Thoughts on The Distribution of Earnings," ***Oxford Economics Papers, 3: 135-146.***

Rubin, Donald. 1978. "Bayesian Inference for Causal Effects: The Role of Randomization," ***Annals of Statistics, 6 (1): 34-58.***

Rubin, Donald. 1979. "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias In Observational Studies," ***Journal of the American Statistical Association,*** 74 (366): 318-329.

School-to-Work Opportunities Act of 1994, May 4, 1994, Public Law 103-239.

Secretary's Commission on Achieving Necessary Skills. 1991. ***What Work Requires of Schools: A SCANS Report for America*** 2000. Washington, D. C.: U.S. Department of Labor.

Secretary's Commission on Achieving Necessary Skills. ***1992. Learning a Living: A Blueprint for High Performance.*** Washington, D. C.: U.S. Department of Labor.

Sheets, Robert, John Conrath, Rogercarole Rogers and Lonnie Taylor. 1993. "Implementing Youth Apprenticeship and Related Training Programs for a Comprehensive Career Development Training System in Business Management at Schools and McDonald's Corporation."

Silverberg, Marsha K., and Alan M. Hershey. *1995. The Emergence of Tech-Prep at the State and Local Levels.* Princeton, New Jersey: Mathematica Policy Research, Inc.

Smith, J. 1995. "A Comparison of the Earnings Patterns of Two Samples of JTPA Eligibles," unpublished manuscript, University of Chicago.

Smith, J. 1994. "Sampling Frame for the Eligible Non-Participant Sample," unpublished manuscript, University of Chicago.

Stern, D., M. Raby, and C. Dayton. *1992. Career Academies: Partnerships for Restructuring American High Schools.* San Francisco, California: Jossey-Bass.

Stern, D., J. R. Stone III, C. Hopkins, M. McMillion and R. Crain. 1994. *School-Based Enterprise: Productive Learning in American High Schools.* San Francisco, California: Jossey-Bass.

Stern, David, Neal Finkelstein, James R. Stone III, John Latting, and Carolyn Dornstife. 1995. *Research on School-to- Work Transition Programs in the United States.* The Stanford Series on Education & Public Policy. Washington, D. C.: Bristol, Taylor & Francis, Incorporated.

Stern, D. et al. *1995. School to Work Research on Programs in the United States.* London, England: The Falmer Press.

Stromsdorfer, Ernst et al. 1985. *Recommendations of the Job Training Longitudinal Survey Research Advisory Panel.* Washington, D. C.: U.S. Department of Labor.

Sum, Andrew M., and Joanna Heliotis. 1993. "Declining Real Wages of Youth." *Workforce, 2,* (2).

Tennessee State Department of Education. 1990. *The State of Tennessee's Student/Teacher Achievement Ratio (STAR) Project Technical Report 1985-1990.*

Texas State Occupational Information Coordinating Committee. 1994. *Automated Student and Adult Learner Follow-up System.* Austin, Texas.

Todd, P. 1995. "Semiparametric Least Squares Estimation of Binary Choice Models with Choice Based Samples Using Local Linear Regression," unpublished manuscript, University of Chicago.

U.S. Congress. *1994. Conference Report to Accompany H. R. 2884, School-to- Work Opportunities Act of* 1994. Washington, D. C.: U. S. Government Printing Office.

U.S. Congress, Office of Technology Assessment. *1995. Learning to Work Making the Transition from School to Work.* Washington, D. C.: U. S. Government Printing Office.

U.S. Congress, House Education and Labor Committee. *1993. Report to Accompany H. R. 2884, School-to-Work Opportunities Act of 1993.* Washington, D. C.: U.S. Government Printing Office.

U.S. Congress, Senate Committee on Labor and Human Resources. 1993. ***School-to- Work Opportunities Act of 1993.*** Washington, D. C.: U.S. Government Printing Office.

U.S. Department of Education. 1991a. ***Combining School and Work Options in High Schools and Two-Year Colleges.*** Washington, D. C.: Office of Vocational and Adult Education.

U.S. Department of Education. 1991b. ***State Comparisons of Education Statistics 1969-70-1993-94.*** Washington, D. C.: National Center for Education Statistics.

U.S. Department of Education. 1994. ***Mini-Digest of Education Statistics, 1994.*** Washington, D.C.: Office of Educational Research and Improvement.

U.S. Department of Education and U.S. Department of Labor. ***1995. School-to-Work Opportunities and the Fair Labor Standards Act.*** Washington, D.C.: National School-to-Work Office.

U.S. Department of Labor. ***1991. What Work Requires of Schools: A SCANS Report for America*** 2000. Washington, D. C.: The Secretary's Commission on Achieving Necessary Skills.

U.S. General Accounting Office. 1991. ***Transition from School to Work Linking Education and Worksite Training.*** Report No. HRD-91-105. Washington, D. C.: U.S. Government Printing Office.

Weiss, Carol Hirschon. 1995. "Nothing as Practical as Good Theory: Exploring Theory-Based Evaluation for Comprehensive Community Initiatives for Children and Families." In New ***Approaches to Evaluating Community Initiatives: Concepts, Methods, and Contexts,*** edited by James P. Connell, Anne C. Kubisch, Lisbeth Schorr, and Carol Weiss. Queenstown, Maryland: The Aspen Institute.

Western Association of Schools and Colleges. 1994. ***Essential Questions and Rubrics for Schoolwide Criteria.***

Whiting, Basil, and Wade Sayer. 1995. "School-to-Work or School-to-What? Exploring Prospects for Building Employer Capacity in School-to-Work Programming." Public/Private Ventures report for the Pew Foundation.

Willis, Robert, and Sherwin Rosen. October 1979. "Education and Self-Selection," ***Journal of Political Economy,*** 87(Supplement): S7-S36.

★ U.S. GOVERNMENT PRINTING OFFICE: 1997 – 418 - 446 I 74652

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

☐ This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☒ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").

EFF-089 (1/2003)